PREDICCIÓN DE LA TENDENCIA CON ÍNDICES PORCENTUALES DE LOS PRECIOS DE BOLSA HORARIOS DEL MERCADO ELÉCTRICO USANDO CLASIFICADORES CON PARÁMETROS ADAPTATIVOS Y VARIAS FUENTES DE INFORMACIÓN

Ing. Ismael Calle Marulanda

Proyecto de grado presentado como requisito parcial
para aspirar al título de Magister en Ingeniería Eléctrica

Director
Mauricio Holguín Londoño, Ph.D.

Co-director
Germán Andrés Holguin Londoño, Ph.D. (c)

Grupo de Investigación en Gestión de
Sistemas Eléctricos, Electrónicos y Automáticos.

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
PROGRAMA DE MAESTRÍA EN INGENIERÍA ELÉCTRICA
PEREIRA
2021

Nota de Aceptación

_____

_____

_____

_____

_____

Firma del Presidente del jurado

_____

Firma del jurado 1 - Evaluador

_____

Firma del jurado 2 - Evaluador

_____

Firma del jurado 3 - Director

Pereira, 02 de Marzo de 2021

Dedico este trabajo de grado a mis padres, Norman Calle y Mery Marulanda, por todo su amor y su esfuerzo.

Agradecimientos

Agradezco principalmente al Creador, a la Fuente Original de todas las cosas, que también nos ha hecho creadores para formar al mundo y contribuir al conocimiento. Así mismo, extiendo mi agradecimiento a la persona que me ha enseñado que la fe no es etiqueta de mediocridad sino que es la medida del bienestar social y del desarrollo humano, a mi líder y guía espiritual, que desde sus momentos más difíciles me sigue motivando. A mis padres que sin importar los obstáculos enseñaron a vencerlos, y con su esfuerzo y dedicación me supieron guiar para luchar por mis sueños, mostrándome que aunque ellos no tuvieron las mismas oportunidades fueron capaces de cambiar a una generación entera, contribuyendo mediante la educación al mejoramiento de la sociedad. Y a todos mis amigos, mis familiares y docentes, de una u de otra forma han contribuido a mi desarrollo como profesional.

Gracias por estar presentes.

CONTENIDO

pág.

LISTA DE TABLAS

LISTA DE FIGURAS

1. INTRODUCCIÓN

Con la actual integración de Generadores Distribuidos (GD) y Autogeneradores a Pequeña Escala (AGPE) en el Sistema de Distribución Local (SDL) colombiano nace la oportunidad de comercializar los excedentes de energía que dichos generadores estarán inyectando a la red eléctrica. Cualquier comercializadora que atienda los mercados regulados puede comprar esta energía sin necesidad de realizar convocatorias públicas, con un precio de venta definido por el Precio Horario en la Bolsa de Energía (PHBE). Toda persona que desee implementar alguno de estos sistemas de generación puede acceder a dicho beneficio de acuerdo al principio de no discriminación, si cumple con las normas que para ello publiquen los entes de regulación, control y vigilancia del mercado eléctrico colombiano, y si cuenta con la aprobación por parte del Operador de Red (OR) para su conexión al SDL. Como es un tema abierto a cualquier clase de público, es normal que los usuarios generadores no cuenten con los conocimientos suficientes en los temas de mercado de energía y sus variables económicas, así como la adecuada disponibilidad de la información sobre los precios de bolsa y otros datos requeridos por los expertos en la materia. De esa forma, es poco probable que sea óptima la remuneración económica obtenida debido a la energía excedentaria que inyecten a la red.

Los precios de bolsa son valores definidos por diferentes mercados financieros, esenciales para la economía de un país, influyentes en la capitalización e interesantes para los inversionistas. En Colombia, existe un mercado mayorista de energía que permite el intercambio de ofertas y demandas, hora a hora, entre comercializadores y generadores, definiendo lo que se conoce como la Bolsa de Energía de Colombia (BEC). En este sistema de información, administrado por la empresa XM, los generadores registran diariamente la cantidad de energía que tienen disponible para una hora correspondiente y su precio de venta, en el mercado a corto plazo. En condiciones normales de operación,

los precios de bolsa se fijan de acuerdo al mayor precio de oferta obtenido; en otras condiciones, los precios de bolsa pueden ser intervenidos por demás variables.

1.1. DEFINICIÓN DEL PROBLEMA

Algunos investigadores afirman que el mercado de valores es dinámico, no lineal, caótico por naturaleza y propenso a gran cantidad de ruido, por lo que es difícil predecir su comportamiento de una forma precisa utilizando solamente valores históricos [1] [2] [3]. Sin embargo, realizar un pronóstico preciso del precio del mercado a corto plazo ha sido un campo de investigación ampliamente atractivo, incluso para los inversionistas [4] [5]. Sugiriendo que el pronóstico del mercado de valores puede tener éxito con el uso de herramientas y técnicas que superen el problema de la incertidumbre, el ruido y la no linealidad de los precios [6]. Esto ha llevado a que se propongan múltiples modelos y algoritmos de predicción, entre los cuales se destacan las Redes Neuronales (NNs, del inglés *Neural Networks*), la Lógica Difusa (FL, del inglés *Fuzzy Logic*), las Máquinas de Soporte Vectorial (SVMs, del inglés *Support Vector Machines*) y algunos otros métodos como los que implementan modelos de Markov [7] [2] [8] [9] [10] [11].

En los pronósticos de precios del mercado tanto a corto como a mediano plazo, las NNs han manifestado un buen rendimiento computacional, con una mayor precisión que otros modelos y una buena capacidad de generalización, ya que pueden aprender sistemas dinámicos fácilmente, haciendo uso de un proceso de reentrenamiento que ajusta sus patrones de datos [10] [12], también se afirma que son más efectivas que algunas topologías convencionales para analizar y pronosticar patrones complejos en las series de tiempo financieras [9] [13]. Sin embargo, los modelos tradicionales enfrentan el problema de no ser paramétricos, su escalabilidad no es óptima para tratar entornos con cambios estructurales en los datos con los que fueron entrenados; y aunque esto

les permite ser menos susceptibles al riesgo de especificación errónea del modelo, en comparación con la mayoría de los modelos paramétricos, siguen siendo tolerantes al ruido, pues al basarse en históricos de datos pueden ser reentrenadas con información incompleta o corrupta, causando sobre-entrenamiento o ineficiencia [14] [8]. Adicionalmente, son herméticas, es decir, se comportan como cajas negras, lo que en estos casos dificulta la comprensión de las causas que en determinado momento hacen fluctuar el precio del mercado [2] [12] [15].

Entre otros métodos que se han utilizado para resolver problemas no lineales, se encuentra la FL, adecuada para el pronóstico del mercado de valores y la gestión de riesgos financieros, ya que permite describir y tratar elementos imprecisos e inciertos presentes en un problema de toma de decisiones, incorporando la incertidumbre en bases de datos y algunas características subjetivas en los modelos [15]. Es una técnica de computación flexible, de naturaleza no lineal, y apta para abordar problemas en cuyo tratamiento se exige de experiencia humana específica, evaluaciones subjetivas y justificaciones adecuadas; a diferencia de las NNs no se comporta como una caja negra y ofrece una visión más clara del modelo [6] [16]. No obstante, la generación de reglas difusas para su implementación a menudo resulta ser un desafío que dificulta la escalabilidad y la flexibilidad del modelo, pues estas reglas se generan a partir de un conjunto histórico de datos que en algunos casos puede no estar disponible o puede modificarse en su estructura, y son definidas antes de que comience el proceso de construcción real [15].

Por otra parte, las SVMs se han utilizado la mayor parte del tiempo en la predicción de precios del mercado, especialmente en la tendencia del precio [17]; demuestran ser superiores a diferentes métodos de clasificación individual al minimizar el umbral del error en la clasificación [9], y superan las limitaciones de algunas NNs, ya que permiten obtener soluciones únicas y globalmente óptimas, pues su entrenamiento es equivalente a resolver un problema de programación cuadrática linealmente restringido, lo que las

hace resistentes al sobreajuste; a diferencia de las NNs, que tienen el riesgo de que el algoritmo quede atascado en los mínimos locales debido a su optimización no lineal por descenso de gradiente [8] [9]. Sin embargo, aunque tienen un rendimiento sobresaliente, este se ve afectado por el número de características, requiriendo mayor tiempo computacional para resolver problemas de gran tamaño [8].

Finalmente, se encuentran los modelos de Markov, que aunque se han utilizado para analizar y predecir fenómenos de series temporales, su uso en el mercado de valores ha sido limitado. No obstante, los estudios realizados en este ámbito han considerado que muchos indicadores observables en el mercado de valores (por ejemplo, los rendimientos de las acciones) solo dependen de los estados ocultos en el mercado, y han sugerido que si se utilizan diferentes fuentes de información o combinación entre varios métodos pueden generar buenos modelos de predicción [7] [18] [15]. Además, algunos investigadores plantean que se puede lograr una mejora del rendimiento si en lugar de formar vectores de observación durante un día completo, se toma el rango completo de valores hora a hora o minuto a minuto [19].

De acuerdo con lo mencionado anteriormente, en el presente proyecto de investigación se desea dar respuesta a la pregunta de si es posible determinar un método que permita encontrar un modelo matemático eficiente, para la predicción a corto plazo de la tendencia del PHBE en Colombia.

1.2. JUSTIFICACIÓN

La correcta predicción del PHBE puede facilitar el mercado de energía eléctrica a los GD y AGPE, incentivando a que muchos más de ellos se integren en el SDL. Dicha integración acarrea importantes beneficios para los usuarios generadores, para el sistema eléctrico y para el medio ambiente. Entre los cuales se pueden mencionar: la reducción

de costos de electricidad de los usuarios autogeneradores, la posibilidad de que los GD y AGPE reciban una remuneración económica debido a la energía excedentaria que inyecten en la red, la reducción de pérdidas gracias al consumo local de la energía generada, la reducción de inversión en capacidad de generación y líneas de transmisión del sistema eléctrico, e incluso, la reducción de emisiones de CO2 durante la vida útil de la red [20]. Este modelo, al ser de un mercado intradiario, tiende a reflejar los precios mayoristas fluctuantes, en los precios minoristas para los usuarios finales, donde el precio marginal de energía aumenta por la cantidad total consumida [21]. En consecuencia, al conocer previamente el PHBE, los autogeneradores también pueden ser motivados a la Desconexión Voluntaria por Demanda (DVP), utilizando los aparatos domésticos de alta carga en las horas de poca actividad de consumo; lo cual les permite incrementar la cantidad de energía excedentaria que inyectan a la red, disminuir sus costos de electricidad, contribuir a la estabilidad del sistema eléctrico al reducir la relación pico-promedio resultante en la demanda de carga, y generar una potencial caída en los niveles de emisión de SO2 y NOx [21]. Sin embargo, dada la posibilidad de que en un futuro cercano, más GD y APGE se integren en el SDL, se hace necesario enfrentarse a la incertidumbre de un modelo cambiante con tendencias que prácticamente no se conocen; donde el mercado eléctrico colombiano, que mayormente depende de las reservas hídricas, seguramente sufrirá modificaciones.

Algunas investigaciones sugieren que las Artificiales (ANNs, del inglés *Artificial NNs*) son una herramienta simple, potente y flexible para pronosticar valores del mercado; proporcionan una mejor solución para modelar relaciones no lineales [22]. Además, son comunmente utilizadas para la predicción de precios en mercados financieros [4]. Pero a pesar de varios modelos presentados, han requerido de tareas complejas a la hora de mitigar el error [14] [12]. Por otro lado, las NNs Recurrentes (RNNs, del inglés *Recurrent NNs*), han demostrado que la predicción de precios de bolsa a corto plazo, haciendo uso

de un histórico de datos intradiarios, es adecuada para los mercados de energía desregulados, debido a la optimización de cálculo computacional que presentan; sin embargo, tienen la desventaja de que su fase de entrenamiento es lenta, dependiendo del tamaño de los datos y del número de parámetros de la red [22]. En otras investigaciones, las NNs Profundas (DNNs, del inglés *Deep NNs*) han manifestado un mejor funcionamiento que los modelos autorregresivos lineales, en el conjunto de entrenamiento, pero la ventaja desaparece principalmente en el conjunto de prueba [10], y aunque su predicción es más precisa que en las RNNs, son redes complicadas y utilizas una gran cantidad de parámetros que hacen que la varianza aumente [14].

Otros métodos como las SVMs han demostrado tener mejor precisión y rendimiento que algunas NNs y que los razonamientos basados en casos (CBR, del inglés *Case-based Reasoning*), para problemas de predicción de tendencia en precios del mercado, siendo una herramienta prometedora [9]. Sin embargo, pueden ser superadas en precisión por otros clasificadores, como Bosques Aleatorios (RF, del inglés *Random Forests*), clasificadores ANN y XGBoost [17], por lo que se recomienda analizar varias fuentes de información para proporcionar indicadores más sólidos que la correlación del precio histórico [11] [23], o combinarlas con otras técnicas para desarrollar un modelo de pronóstico eficiente [8]. Lo mismo se ha hecho con los modelos de Markov y la FL, que al combinarse, logran mayor precisión en la predicción que las ANNs tradicionales, aunque presentan mucha complejidad computacional en algunos casos [15] [19] [7] [18].

Cabe mencionar que es necesario el desarrollo de metodologías que permitan la predicción de precios de bolsa, haciendo uso de parámetros adaptativos, dada la incertidumbre en la tendencia del mercado. Además, se hace necesario sugerir que la predicción de la tendencia del precio, en lugar del valor, puede reducir el problema de la no linealidad presentada en este tipo de mercado, para obtener un modelo con eficiencia y precisión sobresalientes; teniendo en cuenta que, al ser el precio de bolsa ampliamente atractivo

para los inversionistas, el uso de índices porcentuales en la tendencia puede favorecer la predicción, identificando la diferencia porcentual respecto al precio anterior con la que tiende a cambiar el precio siguiente.

Por todo lo anterior, en este trabajo de grado se propone una metodología que permite predecir, a corto plazo, la tendencia con índices porcentuales del PHBE en el mercado eléctrico colombiano, haciendo uso de clasificadores con parámetros adaptativos, y basándose en dos o más fuentes de información; mitigando así el problema de no linealidad, escalabilidad y parametrización que se puede presentar al momento de modificar el modelo del mercado eléctrico, y generando indicadores sólidos que permitan una buena calidad de predicción.

1.3. OBJETIVOS

1.3.1. Objetivo General

Desarrollar una metodología que permita predecir, a corto plazo, la tendencia con índices porcentuales del PHBE en el mercado eléctrico colombiano, haciendo uso de clasificadores con parámetros adaptativos para series de tiempo financieras, como es el caso de los HMMs combinados con las ANNs; y basándose en dos o más fuentes de información, relacionadas con el consumo de energía y con algunas variables económicas.

1.3.2. Objetivos Específicos

1. Realizar un análisis de la información documentada en el estado del arte con el fin de aplicar clasificadores con parámetros adaptativos para series de tiempo financieras, como es el caso de los HMMs combinados con las ANNs.

2. Definir las características, índices porcentuales y parámetros primordiales para la

clasificación de la tendencia de los precios de bolsa con base en registros históricos y varias fuentes de información.

3. Desarrollar una metodología de clasificación de características con parámetros adaptativos e índices porcentuales para la predicción de la tendencia del PHBE.

4. Implementar, comprobar y validar la metodología propuesta con base en la funcionalidad, eficiencia, sostenibilidad y reproducibilidad.

1.4. DECLARACIÓN DE ORIGINALIDAD

Se declara que esta investigación es original, ya que sus contenidos son producto de la directa contribución intelectual del autor. Todos los datos y las referencias bibliográficas están directamente citadas e identificadas, y el software utilizado para la implementación y validación de la metodología ha sido desarrollado personalmente; en los casos que así lo requieran, se cuenta con las debidas autorizaciones de quienes poseen los derechos patrimoniales. Por lo tanto, el autor se hace responsable de cualquier litigio o reclamación relacionada con derechos de propiedad intelectual, exonerando de responsabilidad a la Universidad Tecnológica de Pereira.

Como evidencia de la originalidad se realizan los trabajos derivados, consignados en la sección 3.4.3.

2. MARCO TEÓRICO Y CONCEPTUAL

2.1. ESTADO DEL ARTE

Diferentes tipos de NNs se han implementado para resolver problemas de predicción en los mercados de valores [9]. En el caso de las ANNs, muchas variaciones se han utilizado, como las NNs artificiales de enlace funcional (FLANNs, del inglés *Functional Link Artificial NNs*), que presentan predicciones de precios exitosas [12]. En [3] se utiliza ANN como clasificador haciendo uso de Análisis de Componentes Principales (PCA, del inglés *Principal Component Analysis*) para la predicción de la tendencia del precio. Otros tipos de NNs se han propuesto, como las DNNs de [10], [14] y [24] para la predicción y análisis de precios del mercado; las NNs convolucionales (CNNs, del inglés *Convolutional NNs*), como es el caso de [4]; y las RNNs, como la presentada en [22], para la predicción de precios de bolsa a corto plazo, haciendo uso de un histórico de datos intradiarios. Además, en [2] se implementa una metodología de multifiltros de NNs (MFNN, del inglés *Multi-Filters NNs*) integrando múltiples clases de RNNs y CNNs.

Por otro lado, los modelos basados en SVMs han sido ampliamente utilizados, como es el caso de [23], donde se muestra un modelo SVM con aprendizaje multinúcleo (MKL, del inglés *Multi-kernel Learning*) para las predicciones en mercados de valores, una técnica en la que se combinan las fluctuaciones históricas de los precios del mercado con dos o más fuentes de información, como el volumen de negociación y las noticias. En el análisis del estado del arte de [9] se manifiesta que las SVMs se pueden adaptar en regresiones para predecir valores de series de tiempo financieras, y en ese caso se conocen como Regresión de vectores de soporte (SVR, del inglés *Support Vector Regression*). En otras investigaciones, como [5], utilizan los SVMs junto a otros modelos basándose en otras fuentes de información como el impacto de las emociones y los sentimientos por eventos

locales y globales.

Así mismo, en [16] se propone un modelo de series temporales difusas con buen desempeño en pronóstico y precisión. En [7], [18] y [19] se implementa el Modelo Oculto de Markov (HMM, del inglés *Hidden Markov Model*) como uno de los modelos más usados para los pronósticos en el mercado de valores, ya que se ajusta bien al escenario de la vida real, haciendo uso de varias fuentes de información como las emociones en las redes sociales.

Y por último, en [25] se realiza un análisis comparativo entre nueve modelos de aprendizaje automático (Árbol de Decisiones, Bosque Aleatorio, Refuerzo Adaptativo (Adaboost, del inglés *Adaptive Boosting*), Refuerzo de Gradiente eXtreme (XGBoost, del inglés *eXtreme Gradient Boosting*), Clasificador de Vectores de Soporte (SVC, del inglés *Support Vector Classifier*), Bayes Ingenuo, K-vecinos más cercanos (KNN, del inglés *K-Nearest Neighbors*), Regresión Logística y ANN) y dos potentes métodos de aprendizaje profundo (RNN y Memoria Larga a Corto Plazo (LSTM, del inglés *Long short-term memory*), para la predicción de la tendencia en el mercado de valores.

2.2. SERIES DE TIEMPO FINANCIERAS

Las series de tiempo financieras son una disposición ordenada de datos, en intervalos de tiempo equidistantes, pertenecientes a diversas áreas de la economía; estos intervalos de tiempo pueden ser de cada segundo, minuto, hora, día, semana o incluso cada año. En recientes investigaciones se ha evidenciado un aumento considerable en el análisis y desarrollo de modelos que permitan describir su comportamiento de una forma razonable con el fin de pronosticar la variabilidad de los datos en el futuro, basándose en las observaciones de un histórico de datos; contribuyendo, de esta forma, a la toma de decisiones que puede ser de vital importancia en el mundo financiero y de inversionis-

tas. Es poco común que en las series de tiempo financieras los datos se midan en cada momento o que la serie de tiempo sea continua, como lo son las tomas de temperatura o medidas de caudal; por tanto, las series de tiempo financieras se identifican mayormente con las series de tiempo discretas, donde la toma de datos se realiza en intervalos de tiempo que pueden ser horarios, diarios, semanales, mensuales o anuales. Sin embargo, al considerar la serie de tiempo discreta, se espera que la variable asociada a esta se tome como una variable continua en una escala numérica real. Dichas series de tiempo pueden contener datos relacionados con una sola variable (serie de tiempo univariada) o con más de una variable (serie de tiempo multivariable) [26].

Una serie de tiempo asociada a la variable X, sobre el tiempo establecido T, se denota por $X=\{X_t : t \in T\}$, donde X_t es el valor que toma X en un tiempo t [27]. Por ejemplo, el PHBE donde t cambia en intervalos de hora a hora.

Cualquier serie de tiempo esta caracterizada por un componente de tendencia (T) que representa el resultado de los movimientos a largo plazo y es el componente principal de la serie (puede mostrar un comportamiento ascendente, descendente o constante durante un largo período de tiempo). También puede contener un componente cíclico (C) que se manifiesta en un comportamiento irregular del período, como sucede en la mayoría de las series de tiempo financieras. Por otra parte, algunas series se caracterizan por su componente estacionario (S) que hace que se repitan en intervalos periódicos regulares; y en general, existe otro componente que no se puede predecir, y es el componente de variación irregular o aleatoria (I) que también afecta comúnmente a las series de tiempo financieras [26]. Todos estos componentes se pueden combinar de diferentes maneras, ya sea en una multiplicación o en una suma como se muestra a continuación, donde Y(t) es la observación de la serie de tiempo:

$$Y(t) = T(t) \times C(t) \times S(t) \times I(t)$$

$$Y(t) = T(t) + C(t) + S(t) + I(t)$$

2.2.1. Precio de la Bolsa de Energía en Colombia

El precio de corto plazo de la energía eléctrica en Colombia, es conocido como Precio de Bolsa. Es un valor único para cada hora del día, y se determina por medio de la ejecución de un modelo de despacho de optimización horaria sin restricciones de transmisión, conocido como despacho ideal; considerando de manera especial las características técnicas de los recursos de generación [28]. Corresponde al mayor precio de oferta presentado por las unidades programadas en el despacho ideal que no presentan inflexibilidad; sin embargo, puede sufrir de intervenciones de acuerdo a ciertos procedimientos, teniendo en cuenta lo definido en el "Código de Operación" para las plantas de generación hidroeléctrica con embalse. Este precio, también es limitado por el precio máximo en el que se puede vender la energía en el país (Precio de Escasez), el cual se calcula mensualmente teniendo en cuenta los costos variables asociados al Sistema Interconectado Nacional (SIN) y al precio del combustible [29].

Según lo anterior, la BEC no se diferencia a profundidad con lo que normalmente se hace en los mercados globales, es una figura comercial que cumple su funcionalidad de amortiguador para cubrir las fluctuaciones de la oferta y la demanda, al permitir la compra y venta de energía en un ambiente de competencia. Tiene como objetivo operar el mercado eléctrico nacional de modo que se garantice la mitigación de los costos en la operación del SIN para beneficio de generadores, comercializadores, usuarios industriales y usuarios finales. A su vez, coordina el despacho de los recursos de generación de acuerdo con las ofertas más económicas en libre competencia entre los distintos agentes. En resumen, la BEC presta los siguientes servicios: recepción de ofertas y asignación de los recursos para cubrir la demanda del mercado; despacho económico de los recursos de generación del SIN, tanto térmicos como hidráulicos, mediante ofertas de precios y

cantidades de energía; y suministro de información operativa a todos los agentes del mercado de energía eléctrica para la toma de decisiones.

De esta forma, como también sucede en los mercados de bolsa para cualquier tipo de energía primaria (ejemplo: la bolsa de combustible), un aumento del PHBE representa una disminución en los suministros disponibles, y una caída representa un aumento de suministros [30]. Este comportamiento, sensible a los niveles de consumo de energía eléctrica, se manifiesta de forma dinámica, no lineal, no estacionaria, no paramétrica, ruidosa y caótica, al ser afectado por muchos factores altamente interrelacionados, entre los que se pueden distinguir: las variables económicas (tasas de interés o de cambio, precios de productos básicos), las variables la industria (crecimiento de la producción), las variables empresariales (resultados y rendimientos), las variables psicológicas de los inversores y las variables políticas o normativas [3]. Es por lo anterior que, algunos trabajos como [5], [7] y [31] exploran el impacto de las emociones, en las redes sociales y en otros medios, respecto al movimiento del mercado de valores.

La empresa XM, al ser la encargada de administrar el sistema de información del mercado eléctrico Colombiano, suministra públicamente algunos datos de interés. En la figura 1 se puede apreciar el promedio anual del Precio de Bolsa de Energía desde el año 2000 hasta inicios del año 2021; además, se evidencia que el Precio de Escasez se empezó a registrar desde Diciembre del año 2006, y en los años 2017 al 2020 se activó como limitante del Precio de Bolsa. Cabe resaltar que, conforme a la información de datos históricos encontrada en [32], para el Precio de Bolsa Nacional se evidencia una serie con intervalos de tiempo horario, tal como lo muestra la figura 2 con datos del promedio diario desde el año 2008. Para el Precio de Escasez se evidencia una serie con intervalos de tiempo mensual, con datos históricos diarios desde el año 2008, como se muestra en la figura 3. Estos históricos de datos son actualizados por XM en diferentes intervalos de tiempo que pueden ser variables.

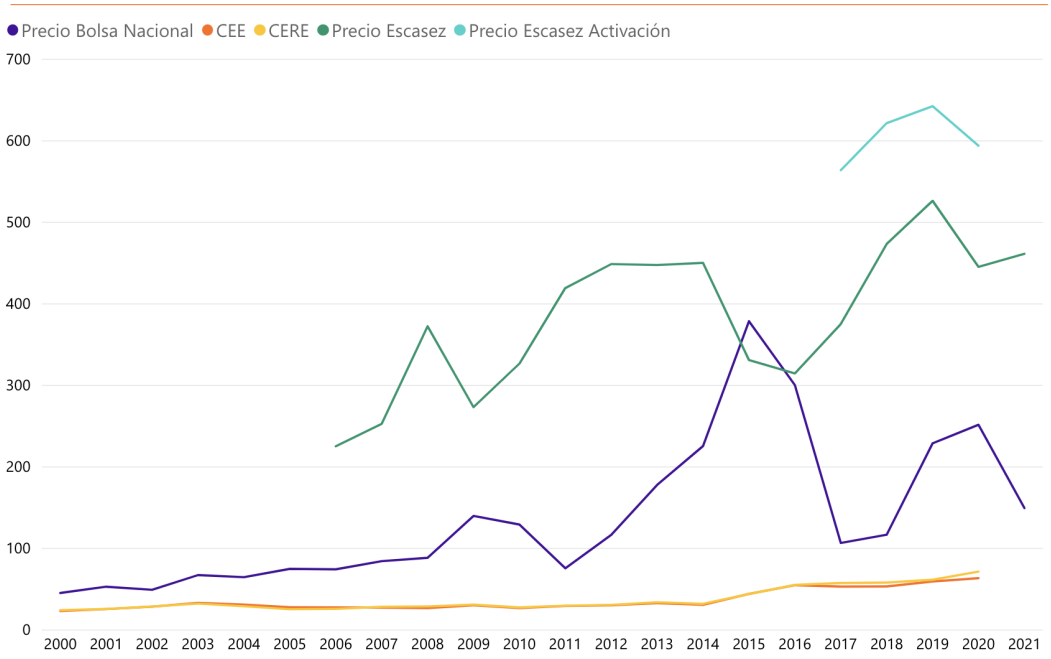


Figura 1. Precio promedio de bolsa COP $ [33]

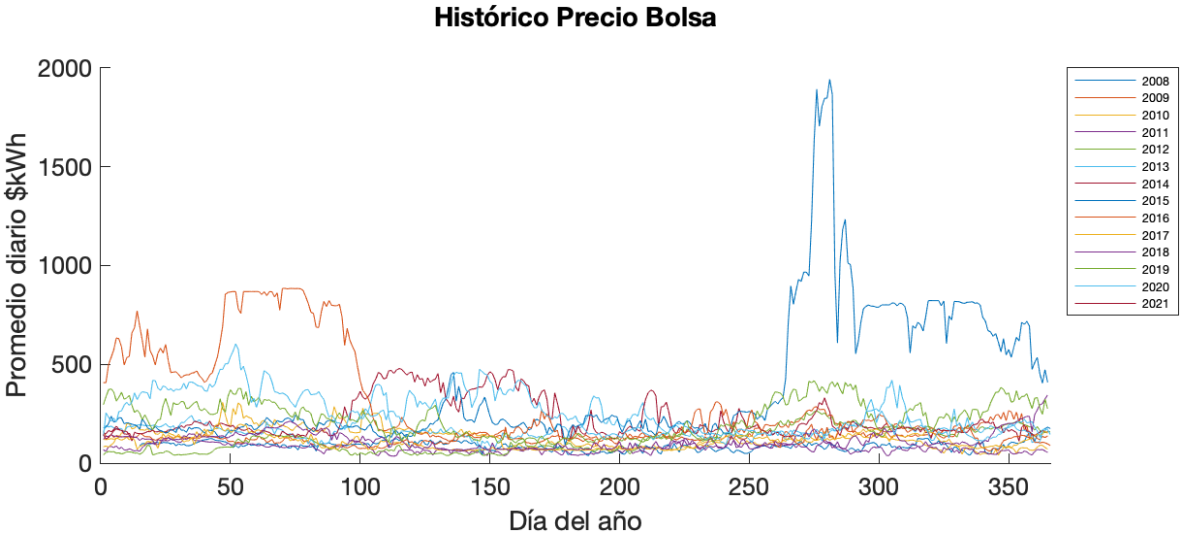


Figura 2. Histórico PHBE. Fuente: autor

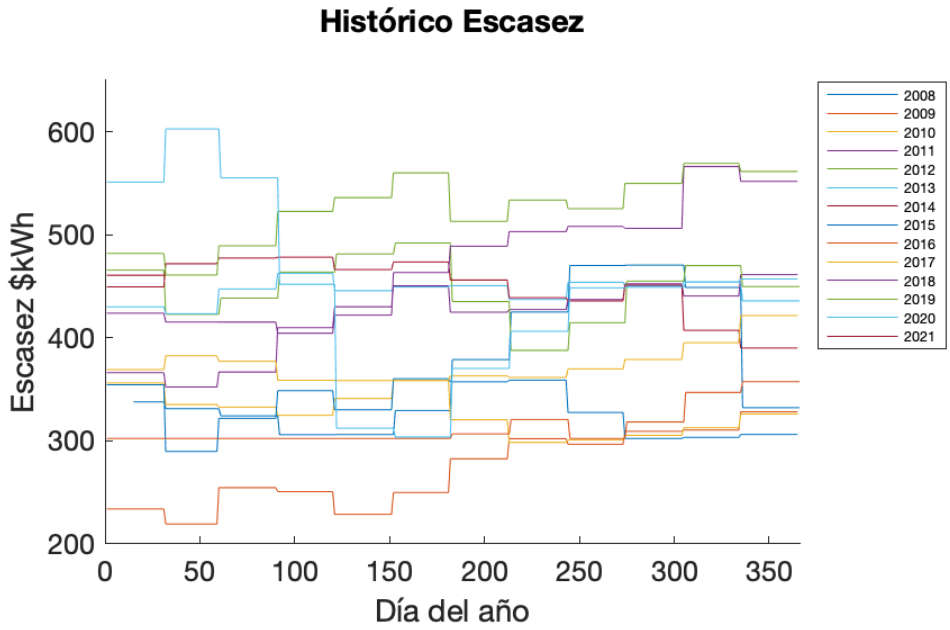


Figura 3. Histórico precio escasez. Fuente: autor

2.2.2. Demanda de energía en Colombia

Dada la relación del PHBE con los suministros disponibles de energía y, consecuentemente, con el consumo de la misma, se hace necesario considerar los históricos de datos que también son suministrados por XM, en [34], como el que se observa en la figura 4 sobre la demanda de energía eléctrica en Colombia desde el año 2008.

La demanda de energía eléctrica en el país es una serie con intervalos de tiempo diario, registrada desde el año 2000, que se utiliza para realizar la asignación de la generación de energía eléctrica teniendo en cuenta otros factores como los precios ofertados por los generadores y algunas restricciones [35]. Se puede apreciar en la figura 5 que esta serie se encuentra marcada por una tendencia que depende del tipo de día. Por ejemplo, entre el lunes y el viernes, sin incluir días festivos, se mantiene un consumo similar. El sábado mantiene un consumo algo particular, mientras que los días domingos y festivos es posible agruparlos dado su consumo de energía eléctrica diferente a los otros días.

Adicionalmente, existen 3 puntos importantes que caracterizan la curva de demanda de energía, estos son, de 5am a 7am, de 11am a 1pm y de 6pm a 9pm, siendo este último punto el de mayor consumo de potencia eléctrica en el país.

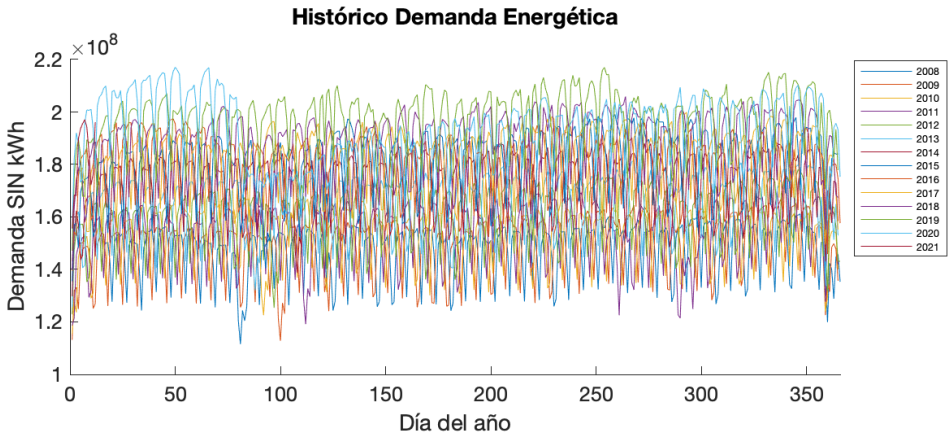


Figura 4. Histórico demanda energía SIN. Fuente: autor

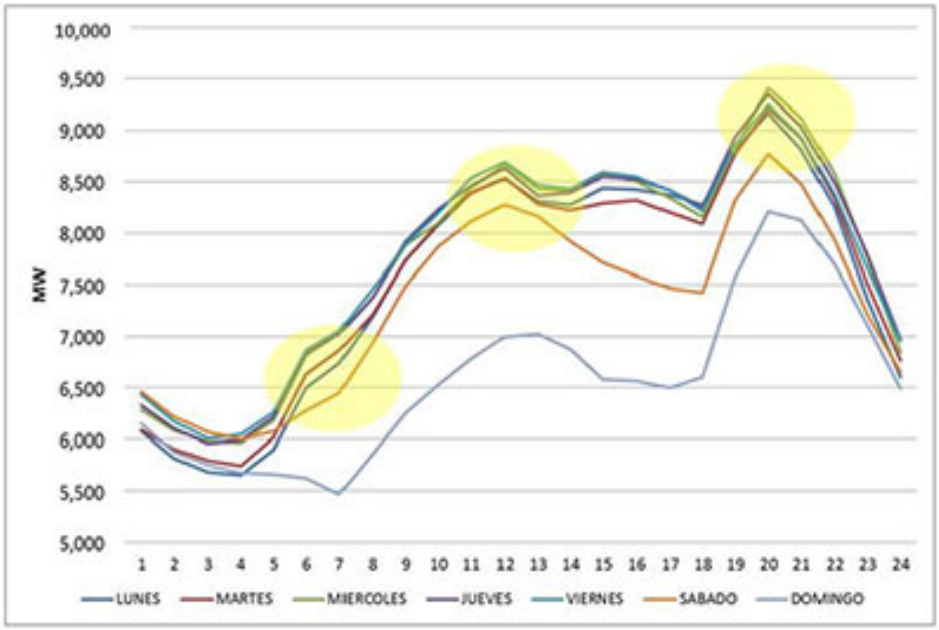


Figura 5. Curva diaria de Demanda de Energía en Colombia [35]

2.2.3. Índice de Capitalización Colombiano (COLCAP)

El COLCAP es un índice de capitalización que refleja las variaciones de los precios de las acciones más líquidas de la Bolsa de Valores de Colombia (BVC), donde la participación de cada acción en el índice está determinada por el correspondiente valor de la capitalización bursátil ajustada (flotante de la compañía multiplicado por el último precio). La canasta del índice COLCAP está compuesta por mínimo 20 acciones de 20 emisores diferentes [36].

Aunque la BVC y la BEC no tienen una relación estrictamente directa, el índice COLCAP se considera importante en este estudio dada la relación del PHBE con las variables psicológicas de los inversores y con el impacto de las emociones. Según la investigación [37], donde se propone una metodología de construcción para el índice de Sentimiento del Inversionista Colombiano (ISIC), el COLCAP se utiliza como variable dependiente del ISIC, este último caracterizado por estar en cinco categorías de medición de sentimiento: miedo extremo, miedo, neutralidad o indiferencia, optimismo y optimismo extremo.

Según los valores históricos encontrados en [36], se cuenta con una serie de tiempo diaria desde el 15 de Enero del año 2008 hasta la fecha, sin embargo, sus registros se aprecian solamente para días hábiles. En la figura 6 se puede apreciar el comportamiento del COLCAP año a año, y en la figura 7 se observa la variación porcentual diaria en los últimos diez años.

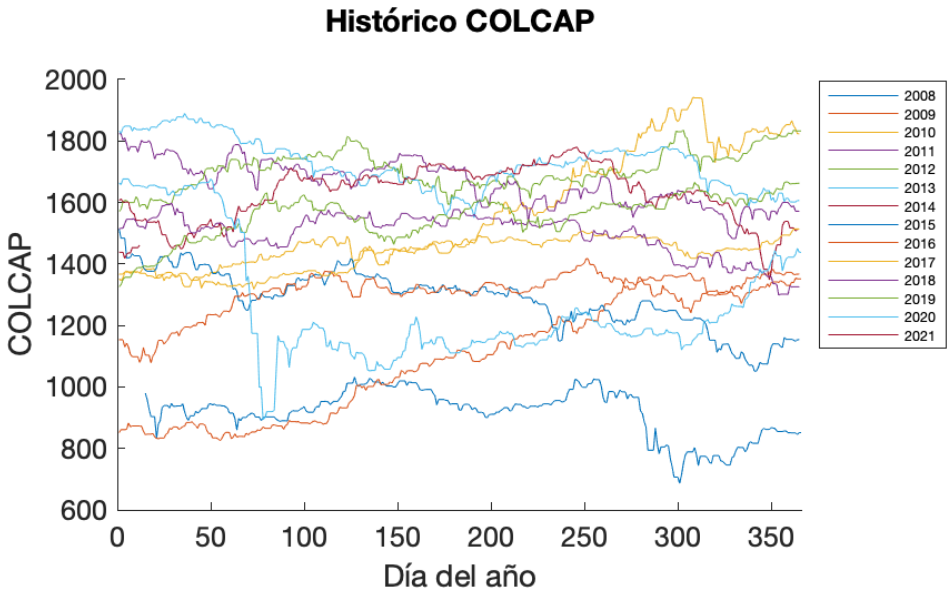


Figura 6. Histórico Valor COLCAP. Fuente: autor

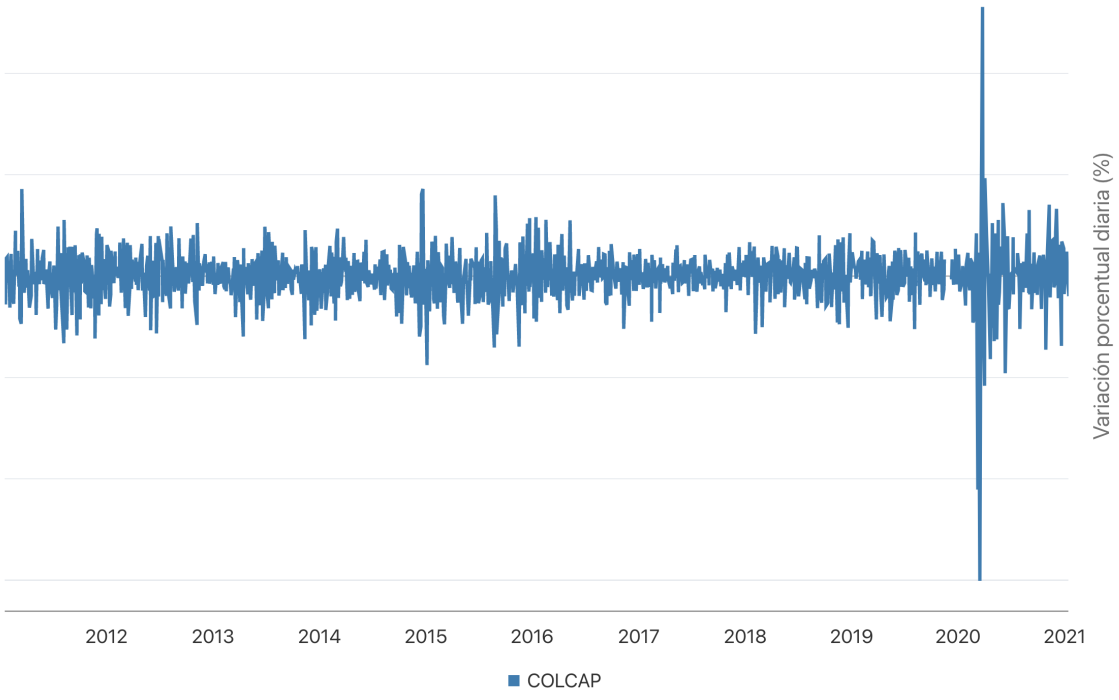


Figura 7. Variación porcentual COLCAP [36]

2.2.4. Generación de energía en Colombia

Colombia cuenta con plantas de generación hidráulica, térmica, eólica, solar y algunos cogeneradores, tanto a gran escala como a pequeña escala. Esta generación puede clasificarse según su actividad principal, en tres tipos [38]:

1. Cogeneración: proceso de producción de energía eléctrica y energía térmica destinada al autoconsumo, o al consumo externo, para sistemas industriales o comerciales. Su principal actividad no es la producción de energía eléctrica.

2. Autogeneración: proceso de producción de energía eléctrica que entrega sus excedentes al SIN luego de atender su consumo propio.

3. Generación: proceso de producción cuya actividad principal es la generación de energía eléctrica. En esta clasificación se encuentran las plantas que tienen una capacidad instalada inferior a 20 MW, excluyendo a los autogeneradores y cogeneradores.

Según los históricos de datos suministrados por XM, en [39], se cuenta con una serie de tiempo horaria registrada desde el año 1995, para la generación despachada (que cubre la demanda de energía en el país) por cada tipo de planta de generación. En la figura 8 se observa el promedio anual histórico de la generación despachada centralmente, desde el año 2008, por cada tipo de generador.

Se evidencia que Colombia es un país que depende altamente del recurso hídrico; esto hace que el fenómeno de El Niño, el fenómeno de La Niña y otros escenarios de hidrología crítica afecten el precio de la energía. Durante los periodos de normalidad hidrológica, la generación hidráulica está en capacidad de abastecer gran porcentaje de la demanda; en contraste con periodos afectados por el fenómeno de El Niño, o periodos secos, las

fuentes de generación como la térmica deben cubrir el porcentaje de la demanda que las fuentes hidráulicas no pueden cubrir, incurriendo en diferentes costos de generación que pueden afectar el PHBE. Con la integración de GD y AGPE en el SDL se espera que el SIN sea suministrado por diferentes tipos de energía en la que la energía solar ha de jugar un papel importante. Como se aprecia en los últimos años.

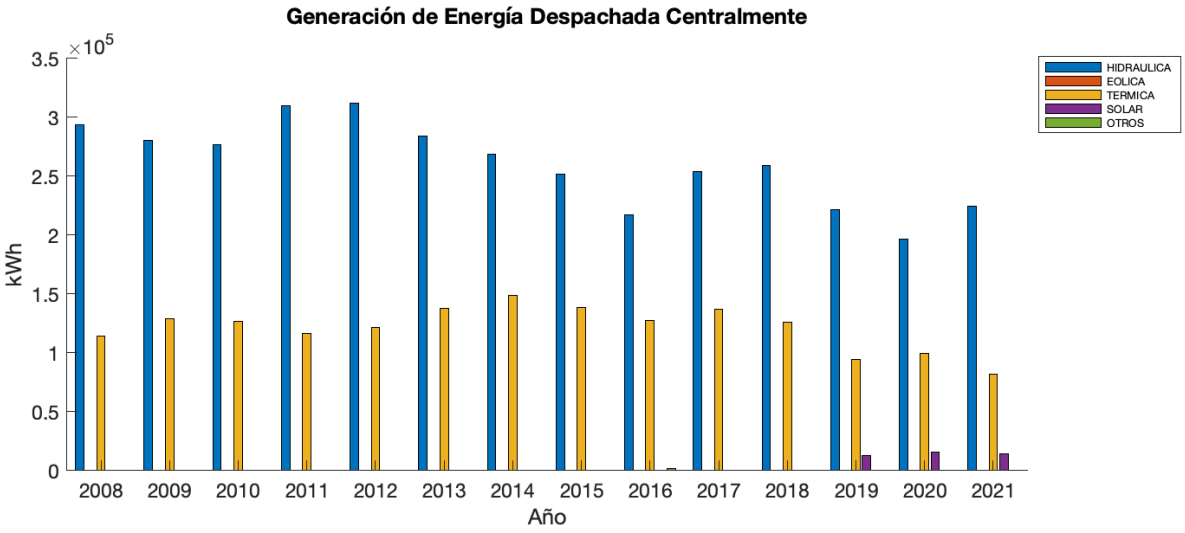


Figura 8. Histórico anual de la generación de energía en Colombia. Fuente: autor

Para el año 2020 se puede apreciar que la generación hidráulica cubrió un 63 % de la demanda, mientras que la térmica y la solar un 32 % y un 5 % respectivamente. Como lo indica la figura 9.

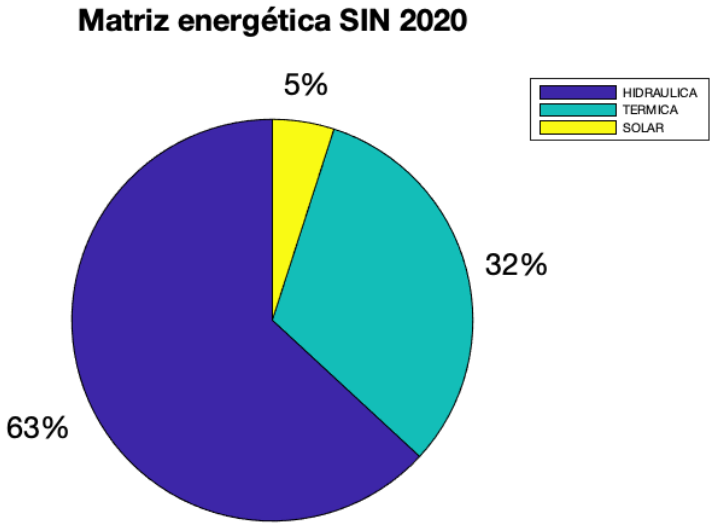


Figura 9. Matríz energética Generación Despachada 2020. Fuente: autor

2.3. CLASIFICADORES UNIVARIABLES POR ANNs

Las ANNs son modelos computacionales que simulan el sistema nervioso de los seres vivos. Pueden definirse como un conjunto de unidades de procesamiento, representadas por neuronas artificiales, interconectadas entre muchos vectores y matrices. Entre sus características más relevantes se encuentra la capacidad de aprendizaje para extraer la relación existente entre las diversas variables de la red; la capacidad de generalización del conocimiento adquirido para la estimación de soluciones; la organización de los datos, permitiendo la agrupación de patrones con características comunes; y tolerancia a fallos gracias al alto número de interconexiones entre las neuronas. Generalmente, los modelos ANNs se conforman por tres capas, como se aprecia en la figura 10: la capa de entrada que es la encargada de recibir los datos de información; las capas ocultas, intermedias o invisibles, compuesta por neuronas que se encargan de extraer los patrones asociados al sistema que se analiza; y la capa de salida, que también está compuesta de neuronas, responsable de producir y presentar las salidas finales de red [40].

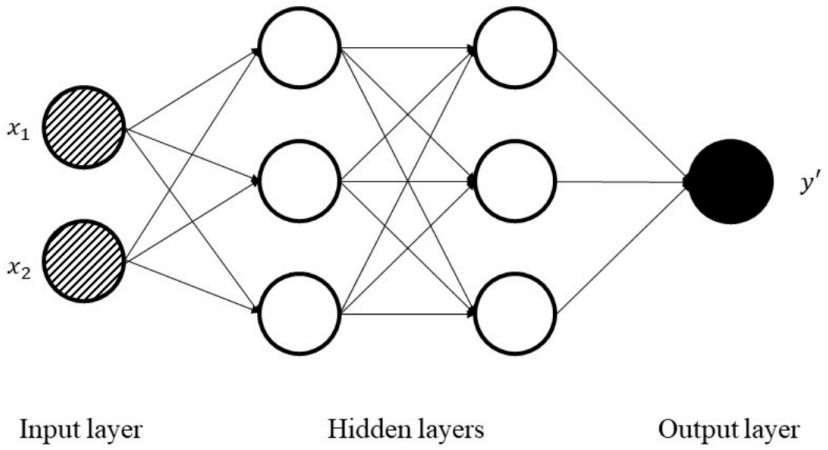


Figura 10. Ilustración esquemática de una ANN [41]

Así mismo, las arquitecturas principales de las ANNs se pueden dividir en cuatro tipos, teniendo en cuenta la disposición e interconexión de las neuronas y la composición de las capas; estos tipos de arquitectura son: redes de alimentación de una sola capa, redes de alimentación de múltiples capas, redes recurrentes y redes de malla [40].

Para que un sistema ANN sea eficiente, usualmente se deben ajustar los datos de entrenamiento, con el fin de lograr un resultado de predicción o clasificación preciso y confiable; estos datos, generalmente se normalizan dentro de los valores límite producidos por las funciones de activación. También, se deben utilizar datos de validación para determinar cuándo detener la fase de entrenamiento en función de una regla de detención temprana, con el fin de evitar el sobreajuste y mejorar la generalización. Y por último, los datos de prueba se deben ingresar al ANN capacitado para proporcionar una medida independiente del rendimiento de la red [3]. La figura 11 representa cada uno de los nodos ocultos; mientras que un nodo toma la suma ponderada de las entradas y lo pasa a través de una función de activación (generalmente una función no lineal), el resultado es la salida del nodo que se convierte en otra entrada de nodo para la siguiente capa; este procedimiento se mueve de la entrada a la salida para cada nodo

con el fin de entrenar la red neuronal.

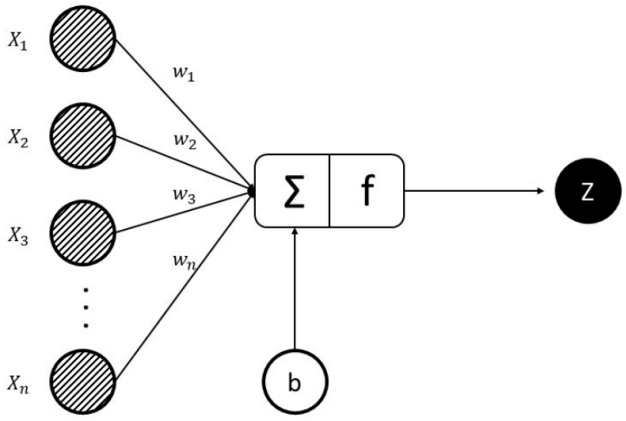


Figura 11. Relación entre entradas y salidas de una ANN [41]

Esta suma se puede suponer como un vector, donde n es el número de entradas para el nodo final; f es la función de activación; X_1, X_2,... y X_n son entradas; w_1, w_2,... y w_n son pesos; y Z es la salida final, tal como se muestra en la ecuación 1 [25].

$$Z = f(x.w + b) = f(\sum_{i=1}^{n} x_i.w_i + b) \tag{1}$$

2.4. CLASIFICADORES MULTIVARIABLES POR HMMs

Los HMMs son modelos estadísticos en los que se analizan procesos de Markov no observados o sistemas de parámetros desconocidos, con el fin de determinar los estados ocultos entre los cuales pueden ocurrir cambios o transiciones, a partir del conjunto de datos que generan una observación. Proporcionan flexibilidad para modelar sistemas de series de tiempo univariadas y multivariadas; especialmente para series de valores discretos, incluidas series categóricas y series de recuentos [42] [19].

Un HMM simple se representa por la notación $\{O_t : t \in N\}$. Si se define $\mathbf{O}^{(t)}$ y $\mathbf{S}^{(t)}$ como una serie de tiempo, del tiempo 1 al tiempo t, se puede resumir un modelo de

este tipo por:

$$P(S_{t+1}|\mathbf{S}^{(t)}) = P(S_{t+1}|S_t), t = 1, 2, 3, ...$$
$$P(O_{t+1}|\mathbf{O}^{(t)}, \mathbf{S}^{(t)}) = P(O_t|S_t), t \in N \tag{2}$$

El anterior modelo, en primer lugar, consta de un proceso de parámetros no observados o número finito de estados $\{S_t : t = 1, 2, ...\}$, en segundo lugar, de un proceso dependiente del estado u observaciones $\{O_t : t = 1, 2, ...\}$, y en tercer lugar, de las probabilidades iniciales de transición de un estado S_i a un estado S_j $\{P_{ij} : i, j = 1, 2, ...\}$. De modo que, cuando se conoce S_t, como estado actual, su información es la que determina la probabilidad de distribución de los estados futuros para O_t, y no los estados u observaciones anteriores, satisfaciendo así la propiedad de Markov. Sin embargo, los HMMs no siempre satisfacen dicha propiedad, por ejemplo, un Bernoulli-HMM de dos estados puede degenerar de manera obvia a la cadena subyacente de Markov, logrando identificar a cada uno de los dos valores observables con uno de los dos estados subyacentes [42].

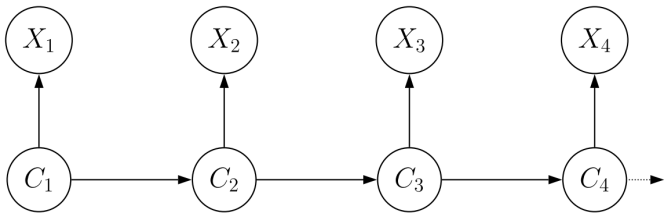


Figura 12. Ilustración esquemática de un HMM básico [42]

Si se supone que el proceso Markoviano es de tiempo discreto y de estado finito observado, y que posee un ruido sin memoria, es decir, hay una secuencia de observaciones O_t que dependen probabilísticamente solo de la transición a_t para el tiempo t. Se puede apreciar $P(O|S)$ de la siguiente manera [43]:

$$P(O|S) = P(O|A) = \prod_{t=0}^{T-1} P(o_t|a_t) \tag{3}$$

Inicialmente se asume que el proceso comienza desde el tiempo 0 hasta el tiempo T, y que los estados s_0 y s_T son conocidos. A es la matriz de transición entre los pares de estados (S_{t+1}, S_t) para cada $P(S_{t+1}|S_t) \neq 0$.

En estos modelos, los estados no son directamente visibles pero las observaciones que dependen del estado sí son visibles. Cada uno de los estados del HMM tiene una distribución de probabilidad sobre las posibles observaciones de salida. Por tanto, la secuencia de observaciones, que dicho modelo genera, proporciona información sobre la correspondiente secuencia de estados. En este caso, el observador no sabe en qué estado puede estar el sistema pero sí tiene una idea probabilística de dónde debería estar [44].

Algunos algoritmos, como el algoritmo de Viterbi, se encargan de encontrar la secuencia de estados S para la cual la probabilidad a posteriori $P(O|S)$ es máxima, y a su vez la secuencia de transiciones A para la cual $P(A|S)$ es máxima, con el fin de otorgar entonces la ruta más corta de estados para la secuencia de observaciones dada. Esto se resuelve encontrando el camino cuya longitud de Probabilidad de Logaritmo (LogLik, del inglés *Logaritm Likelihood*) $-\ln(P(O,S))$ es mínima, ya que $-\ln(P(O,S))$ es una función monótona de $P(O,S)$ y existe una correspondencia biunívoca entre caminos y secuencias [43].

En cuanto a la selección del número de estados ocultos, puede variar de acuerdo con las necesidades de la investigación. Algunos estudios sugieren tres estados ocultos para representar el mercado de valores: estado estable, incierto y de colapso; a los que se les calcula la matriz de transición luego de aplicar el algoritmo Bayesiano de Monte Carlo (haciendo uso de la longitud LogLik) para estimar los parámetros que aumenten la probabilidad de las variables observables [7].

2.5. MÉTODO DE AGRUPACIÓN DE DATOS *K-MEANS*

K-means es un método de agrupación de datos que tiene como objetivo principal separar un conjunto de observaciones en diferentes grupos o *clusters* de acuerdo con la distancia euclideana entre cada observación con el centro del grupo (centroide). Separando de esta forma las observaciones por su media más cercana [45].

El algoritmo consiste en definir un número de *clusters* ($C = \{c_k, k = 1, ..., K\}$) para el agrupamiento de un conjunto de puntos ($X = \{x_i\}, i = 1, ..., n$), mediante la búsqueda de particiones tal que el error al cuadrado entre la media empírica de un grupo y los puntos en el grupo sea mínima. Si μ_k es la media del grupo c_k, el error al cuadrado entre μ_k y los puntos del *cluster* c_k está definido como:

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \tag{4}$$

El objetivo de *K-means* es minimizar la suma del error al cuadrado sobre todos los K grupos

$$J(C) = \sum_{k=1}^{K} J(c_k) \tag{5}$$

Los pasos principales de este algoritmo son los siguientes:

1. Seleccionar el número de *clusters*

2. Calcular la distancia entre todos los puntos al centro del *cluster*.

3. Asociar cada punto al *cluster* más cercano.

4. Recalcular el centro de los *clusters* a partir de los puntos que lo componen.

5. Repetir los pasos hasta que el algoritmo converja o se cumpla el número máximo de iteraciones. El algoritmo converge cuando $J(C) = 0$.

La figura 13 muestra una ilustración del algoritmo *K-means* en un conjunto de datos bidimensional con tres *clusters*.

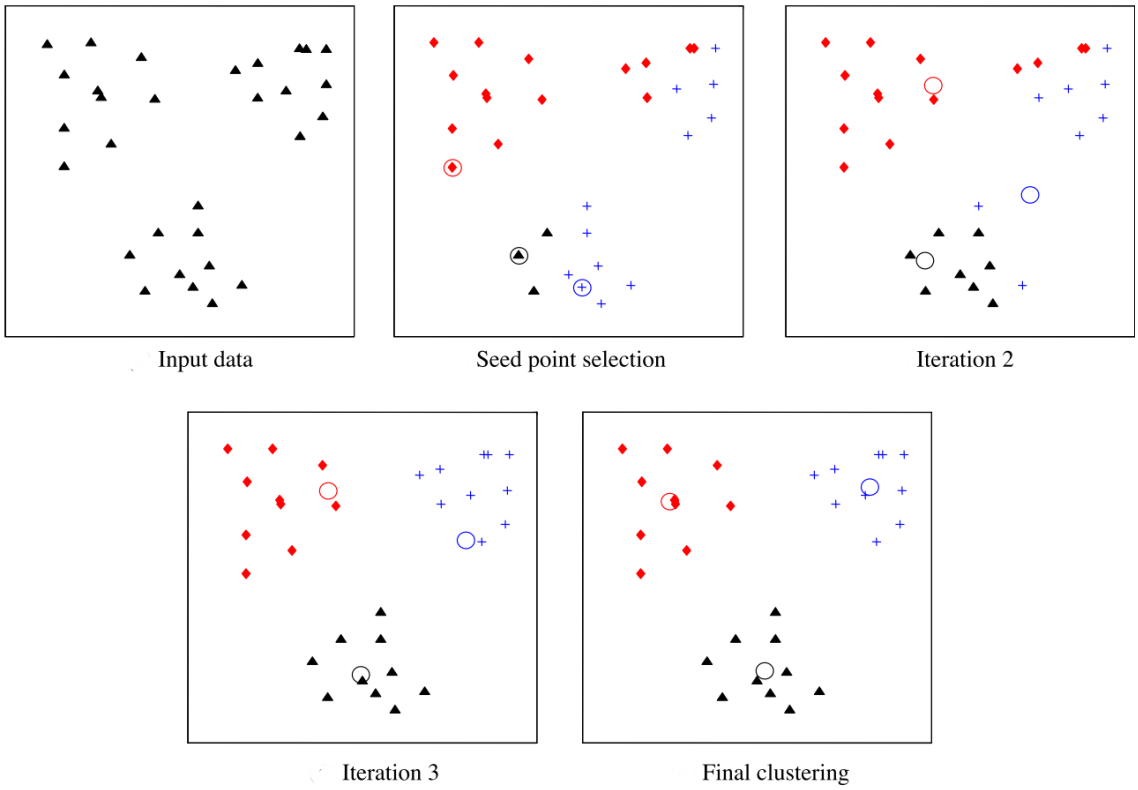


Figura 13. Ilustración del algoritmo *k-means* [45]

2.6. MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS

2.6.1. PCA

El PCA es un método estadístico que extrae un número reducido de factores, o componentes principales, de elementos altamente correlacionados en las variables originales. Tradicionalmente es el método más simple y más eficiente que se usa junto con las

ANNs y en algunos casos con la SVMs para reducir la complejidad computacional de problemas de clasificación [9].

Estos componentes principales pueden ser expresados de la siguiente manera:

$$\begin{cases} Y_1 = A_{11}.X_1 + A_{12}.X_2 + ... + A_{1n}.X_n, \\ Y_2 = A_{21}.X_1 + A_{22}.X_2 + ... + A_{2n}.X_n, \\ \qquad\qquad ... \\ Y_n = A_{n1}.X_1 + A_{n2}.X_2 + ... + A_{nn}.X_n \end{cases} \tag{6}$$

Donde X_i es la variable original, Y_i es el componente principal y A_i es el vector de coeficientes que puede ser estimado maximizando $Var(Y_i)$ con las condiciones de restricción de $A_i^T.A_i = 1$ y $Cov(Y_i, Y_j) = A_i^T.\sum A_i = 0, j = 0, 1, 2, ..., i-1$ [46].

Con base en el vector de coeficientes se puede estimar el peso y el índice de las características más relevantes en un conjunto de datos según su nivel de variabilidad. Entre más variabilidad tenga una característica, puede aportar más información al sistema.

2.6.2. ReliefF

El algoritmo principal de Relief es usado para problemas de clasificación que contienen solamente dos clases, su idea original es que las características de alta calidad deben tener diferentes valores para instancias de diferentes clases y valores similares para instancias de la misma clase. ReliefF nace como una extensión de la familia Relief para ser un poco más robusto y para manejar datos incompletos y ruidosos sin limitarse a los problemas de dos clases. Se utiliza generalmente para la selección de características, basándose en vecindarios, con el fin de reducir la complejidad computacional de un problema [47]. El pseudocódigo que lo representa se puede apreciar a continuación [48]:

Entradas: para cada instancia de entrenamiento un vector de valores de atributo y el valor de la clase.

Salida: el vector W de estimaciones de las cualidades de los atributos.

1. Poner todos los pesos $W[A] := 0.0$;
2. **for** i :=1 **to** m **do begin**
3. seleccionar aleatoriamente una instancia R_i;
4. encontrar k coincidencias H_j más cercanas;
5. **foreach** clase $C \neq clase(R_i)$ **do**
6. desde la clase C encontrar k fallas $M_j(C)$ más cercanas;
7. **for** A :=1 **to** a **do**
8. $W[A] := W[A] - \sum_{j=1}^{k} diff(A, R_i, H_j)/(m.k) +$
   $\sum_{C \neq clase(R_i)} [\frac{P(C)}{1-P(clase(R_i))} \sum_{j=1}^{k} diff(A, R_i, M_j(C))]/(m.k)$;
9. **end**

2.7. VALIDACIÓN CRUZADA

Es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar la independencia entre los datos de entrenamiento y los datos de prueba. Consiste en repetir y recalcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones de los datos. Se utiliza en entornos donde el objetivo principal es la predicción y se requiere estimar la precisión de un modelo que se lleva a la práctica.

Uno de los métodos más comunes de validación cruzada es el de K iteraciones; en este método se dividen los datos de entrada en K subconjuntos de datos, luego se entrenan los

modelos en K-1 subconjuntos para, posteriormente, evaluar el modelo en el subconjunto de datos que no se ha utilizado para el entrenamiento. Este mismo proceso se repite K veces, eligiendo un subconjunto diferente de datos que sea exclusivo para la etapa de prueba o evaluación, con el fin de evitar el sobre ajuste y obtener un modelamiento adecuado [49].

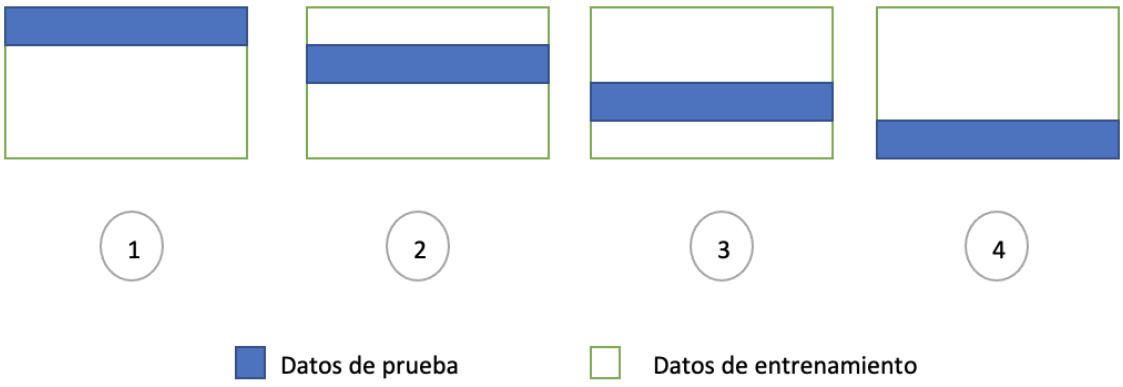


Figura 14. Ejemplo Validación Cruzada. Fuente: autor

En la figura 14 se aprecia un ejemplo de validación cruzada por K iteraciones, donde K toma un valor de 4. Para cada uno de las iteraciones se elige un subconjunto diferente del 25 % para la evaluación y un subconjunto del 75 % para el entrenamiento.

Una vez que se realiza la validación cruzada entre los modelos, se puede ajustar la configuración del modelamiento final de acuerdo a los estándares de rendimiento.

3. METODOLOGÍA Y RESULTADOS

3.1. TRATAMIENTO DE LOS DATOS

La información tomada para este estudio se basa en 5 series de tiempo (PHBE, Generación Despachada Centralmente, Demanda de Energía del SIN, Precio de Escasez y COLCAP) ajustadas a la fecha mínima común entre ellas y a intérvalos de tiempo horario, repitiendo los valores que tienen intervalos de tiempo mensual o diario, para cada hora del día; dando como resultado un conjunto de datos con fechas desde el año 2008 hasta principios del 2021. Cabe resaltar que el Precio de Escasez se conoce con anticipación a la fecha, es decir, el Precio de Escasez del mes de febrero se ha de conocer finalizando el mes de enero; para el caso del índice COLCAP, al ser un valor variable que depende de las acciones de la BVC, su valor es promediado diariamente para después ser registrado en el histórico de datos, dicho promedio es conocido al realizar la subasta de cierre del mercado (actualmente a las 4:00pm Hora Estándar de Colombia) [50]; el valor de la Demanda de Energía del SIN, el PHBE y la Generación Despachada Centralmente son actualizados por XM, 1 o 2 días después, con la posibilidad de que sus valores sean reajustados posteriormente.

Algunas de estas series tienen un componente de tendencia altamente marcado, como es el caso de lo observado en la figura 5, donde dicho componente está definido por el tipo de día. Por ejemplo, entre el lunes y viernes se mantiene un consumo promedio, el cual puede ser clasificado como un día ordinario. El sábado mantiene un consumo particular, el cual puede seguir siendo identificado de esta forma, mientras que los días domingos y festivos es posible agruparlos dado su consumo de energía eléctrica similar. Adicionalmente, entre la media noche y las 7 de la mañana el consumo de energía es bajo, entre las 11 de la mañana y la 1 de la tarde el consumo de energía es medio, y

entre las 6 de la tarde y las 9 de la noche se encuentra el punto de mayor consumo de potencia eléctrica en el país.

El comportamiento del PHBE que se observa en la figura 2 permite sugerir que el componente de tendencia de la serie puede estar influenciado por el número del año o incluso por el mes del año, con algunos datos atípicos. El Precio de Escasez observado en 3 y el índice COLCAP apreciado en 6 no muestran claramente ningún componente de tendencia. Pero en el caso de la Generación Despachada Centralmente, al depender en su mayoría de factores ambientales, se puede suponer que el componente de tendencia está influenciado por la época del año o por la hora del día dependiendo del tipo de generación; además, con la cantidad de generación de energía por tipo se pueden identificar características de escasez o abundancia del recurso de generación y, por ende, de afectaciones en el PHBE.

En consecuencia, se definen los datos de entrada para el sistema como se especifica en la tabla 1, teniendo en cuenta cada serie de tiempo con su respectiva fecha. Estas fechas están identificadas por la ecuación 7, ecuación 8, ecuación 9, ecuación 10 y ecuación 11, con el fin de que el sistema detecte los posibles componentes de tendencia de las diferentes series de tiempo. Además, se toman los datos de generación de energía promediados hora a hora según el tipo de generación (hidráulica, eólica, térmica, solar y otros), como parte de los parámetros adaptativos a la tendencia del mercado eléctrico y a la matriz energética del país. Finalmente, se adiciona el día, mes, año y la hora del dato que se espera predecir.

$$Y_t = \left[y_1, y_2, ..., y_t \right]^\top, y \in N, 2008 \leq y \leq 2021 \tag{7}$$

Y_t es una serie identificada con su respectivo componente de tendencia e intervalos de tiempo horario.

$$M_t = \left[m_1, m_2, ..., m_t \right]^\top, m \in N, 1 \leq m \leq 12 \tag{8}$$

M_t es una serie con intervalos de tiempo horario, identificada con un componente de tendencia y un componente cíclico.

$$D_t = \left[d_1, d_2, ..., d_t \right]^\top, d \in N, 1 \leq d \leq 31 \tag{9}$$

D_t es una serie con intervalos de tiempo horario que comprende valores entre el 1 y el 31 dependiendo del mes, identificada con un componente de tendencia y un componente cíclico.

$$N_t = \left[n_1, n_2, ..., n_t \right]^\top, n \in N, 1 \leq n \leq 7 \tag{10}$$

N_t es una serie con intervalos de tiempo horario, identificada con un componente de tendencia y un componente cíclico.

$$H_t = \left[h_1, h_2, ..., h_t \right]^\top, h \in N, 1 \leq h \leq 24 \tag{11}$$

H_t es una serie con intervalos de tiempo horario, identificada por su componente de tendencia y su componente cíclico.

Por otra parte, al tener en cuenta los diferentes intervalos de tiempo en los que se conocen los valores de los datos tratados, es necesario definir una variable de desfase para cada serie de tiempo, de acuerdo con la última fecha conocida de sus datos; como se muestra en la ecuación 12. Esta variable también establece el plazo de la predicción y se define como un vector que contiene el valor del desfase en horas respecto a la fecha a predecir.

$$V = \left[v_1, v_2, v_3, v_4, v_5 \right], v \in N \tag{12}$$

Donde v_1 es el desfase para el PHBE, v_2 para la Generación Despachada Centralmente; v_3, v_4 y v_5 para la Demanda de Energía del SIN, el Precio de Escasez y el COLCAP, respectivamente.

De esta forma, cada una de las series de tiempo de la tabla 1 se organizan en matrices, como se muestra a continuación:

$$F_t = \begin{bmatrix} Y_t & M_t & D_t & N_t & H_t \end{bmatrix} \tag{13}$$

$$PHB_t = \begin{bmatrix} PB_t & Y_t & M_t & D_t & N_t & H_t \end{bmatrix} \tag{14}$$

$$G_t = \begin{bmatrix} HG_t & EG_t & TG_t & SG_t & OG_t & Y_t & M_t & D_t & N_t & H_t \end{bmatrix} \tag{15}$$

$$DEM_t = \begin{bmatrix} ED_t & Y_t & M_t & D_t & N_t & H_t \end{bmatrix} \tag{16}$$

$$ES_t = \begin{bmatrix} E_t & Y_t & M_t & D_t & N_t & H_t \end{bmatrix} \tag{17}$$

$$COLC_t = \begin{bmatrix} C_t & Y_t & M_t & D_t & N_t & H_t \end{bmatrix} \tag{18}$$

Donde PB_t y E_t contienen los valores hora a hora del PHBE y el Precio de Escasez, respectivamente; en unidades de \$kWh. Las series HG_t, EG_t, TG_t, SG_t y OG_t contienen los valores de generación de energía despachada centralmente, hora a hora, por cada tipo de generación, en unidades de kWh. La serie DEM_t contiene los valores de la demanda de energía en el país, adaptados a un intervalo de tiempo horario y en unidades de kWh. Y, por último, la serie C_t contiene los valores del índice COLCAP diario adaptados a intervalos horarios.

Asumiendo que se va a predecir el PHBE de la hora t, y que los últimos datos que se conocen para cada una de las variables son de la hora $t - V$. Se puede definir la matriz de datos de entrada como se muestra en la ecuación 19.

$$X_t = \begin{bmatrix} F_t & PHB_{t-v_1} & G_{t-v_2} & DEM_{t-v_3} & ES_{t-v_4} & COLC_{t-v_5} \end{bmatrix} \tag{19}$$

Cabe resaltar que para el entrenamiento del modelo se eliminan las columnas repetidas, especialmente de las fechas si el desfase V coincide entre las variables.

DATOS	SERIES DE TIEMPO
Fecha de predicción (F)	Año (Y) Mes (M) Día del mes (D) Día de la semana (N) Hora (H)
PHBE (PBE)	PB Año (Y) Mes (M) Día del mes (D) Día de la semana (N) Hora (H)
Generación (G)	Hidráulica (HG) Eólica (EG) Térmica (TG) Solar (SG) Otros (OG) Año (Y) Mes (M) Día del mes (D) Día de la semana (N) Hora (H)
Demanda (DEM)	ED Año (Y) Mes (M) Día del mes (D) Día de la semana (N) Hora (H)
Escasez (ES)	E Año (Y) Mes (M) Día del mes (D) Día de la semana (N) Hora (H)
COLCAP ($COLC$)	C Año (Y) Mes (M) Día del mes (D) Día de la semana (N) Hora (H)

Tabla 1. Datos y variables de entrada.

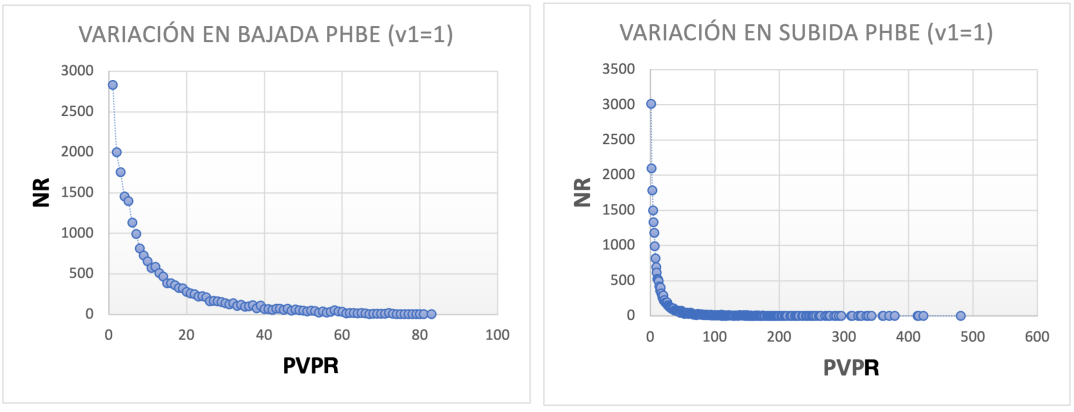
3.1.1. Definición de los índices porcentuales

Teniendo en cuenta el desfase en el que se conocen los datos, se recorren todos los valores de la serie PB_t obteniendo el porcentaje de variación del precio (PVP) entre un dato actual y un dato anterior, como se muestra en la ecuación 20. Luego se redondea al entero más cercano según la ecuación 21, donde $R(x)$ es la función de redondeo.

$$PVP = \left[\frac{PB_t - PB_{t-v_1}}{PB_{t-v_1}} \times 100 \right], PVP \in \mathcal{M}_{n\times 1}(N) \tag{20}$$

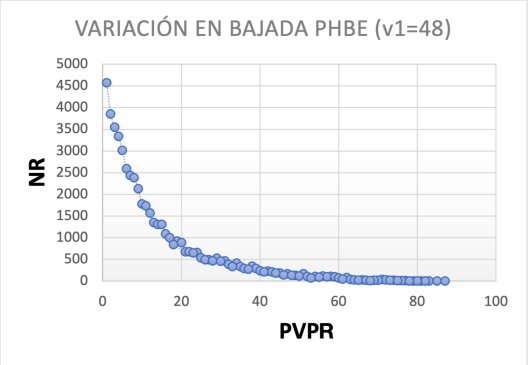
$$PVPR = R(PVP) \tag{21}$$

Seguidamente se procede a hallar el número de repeticiones (NR) para cada valor del $PVPR$, dentro de este mismo conjunto de datos, con el fin de definir los índices de tendencia del precio. Así mismo, se gráfica la dispersión de los datos del $PVPR$ y el NR como se aprecia en las figuras 15, 16, 17, 18 y 19.

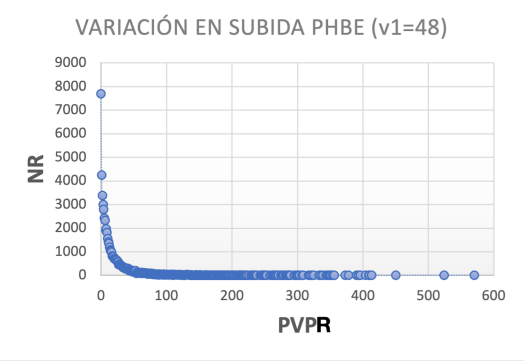


(a) Variación en Bajada (b) Variación en Subida

Figura 15. Variaciones del PHBE con v_1=1h. Fuente: autor

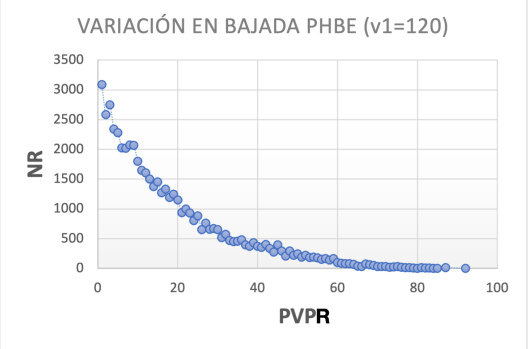


(a) Variación en Bajada

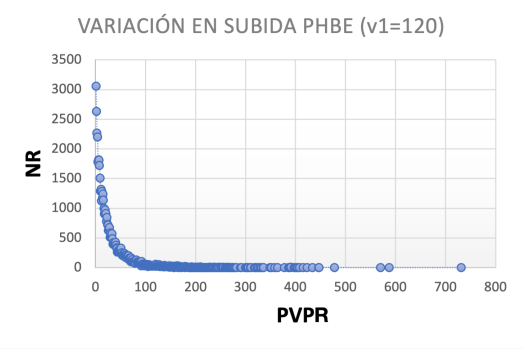


(b) Variación en Subida

Figura 16. Variaciones del PHBE con v_1=48h. Fuente: autor

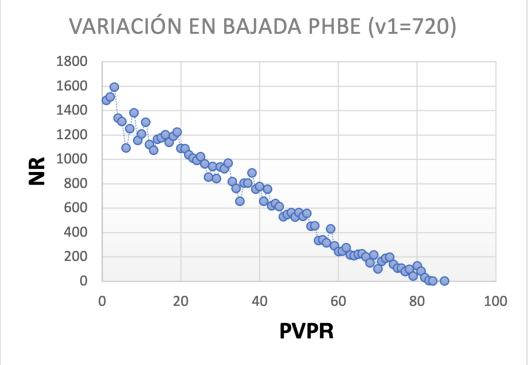


(a) Variación en Bajada

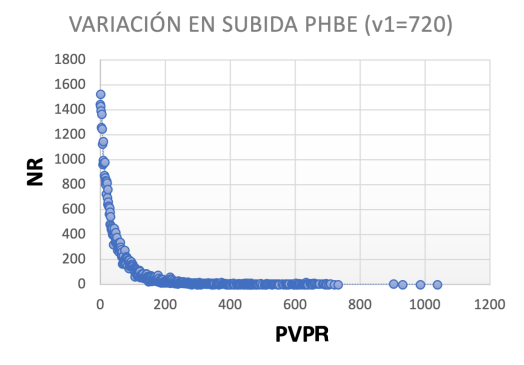


(b) Variación en Subida

Figura 17. Variaciones del PHBE con v_1=120h. Fuente: autor

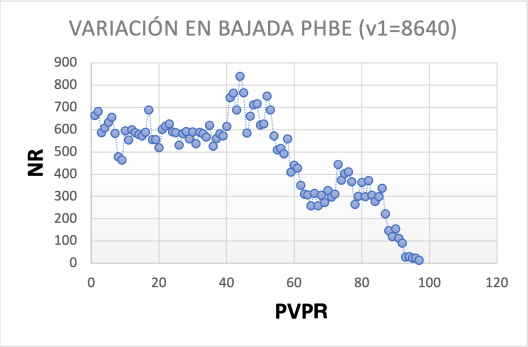


(a) Variación en Bajada

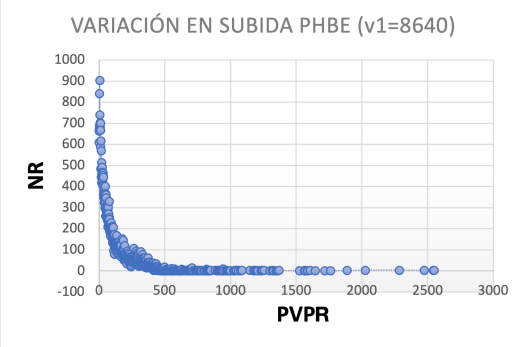


(b) Variación en Subida

Figura 18. Variaciones del PHBE con v_1=720h. Fuente: autor



(a) Variación en Bajada



(b) Variación en Subida

Figura 19. Variaciones del PHBE con v_1=8640h. Fuente: autor

En los casos donde el PHBE sube respecto a un precio anterior, existe una correlación exponencial negativa entre el $PVPR$ y el NR sin importar el desfase v_1 con el que se calcule la diferencia de los precios. Esta correlación se va debilitando a medida que el desfase aumenta, es decir, si se calcula el $PVPR$ y el NR con 1 hora de diferencia (comparando el precio de bolsa en la hora h con el precio de bolsa en la hora h-1) la correlación es muy fuerte, pero si se calculan con un desfase de 8640 horas o 360 días la correlación es débil como se observa en la figura 19b.

De igual manera, en los casos donde el PHBE baja respecto a un precio anterior, existe una correlación exponencial negativa entre el $PVPR$ y el NR. Sin embargo, a medida que el desfase aumenta, la correlación se debilita y tiende a volverse lineal negativa o posteriormente a volverse nula, como se observa en las figuras 18a y 19a.

Por otra parte, se evidencia que cuando el PHBE baja respecto a un precio anterior es poco probable que el porcentaje de diferencia o PVP de bajada sea superior al 100 %, aunque este aumenta levemente de manera proporcional al desfase. Y, cuando el PHBE sube, el PVP tiende a ser mayor que de bajada y aumenta considerablemente de forma proporcional al desfase. Esto quiere decir que es alta la probabilidad de que el PHBE aumente más de lo que disminuya.

$PVPR$ [%]	NR		
	v_1=1	v_1=48	v_1=120
<-5	12664	35694	41770
>=-5 $\wedge$ <=5	87882	41888	29438
>5	13309	36226	42528
TOTAL	113855	113808	113736

Tabla 2. Variación del PHBE.

Para la definición de los índices porcentuales en la variación del PHBE, se establece un desfase máximo de 120 horas con el fin de conservar la correlación entre el $PVPR$ y el NR, y así poder mantener la probabilidad de ocurrencia de un evento de subida o de bajada para cada índice porcentual. En consecuencia, el plazo máximo para la predicción del PHBE en este trabajo se fija en 5 días a partir de la última fecha conocida de los datos.

ÍNDICES PORCENTUALES [%]
<-5 $\wedge$ >-100
>=-5 $\wedge$ <=5
>5 $\wedge$ <800

Tabla 3. Índices porcentuales para tendencia del PHBE.

En la tabla 2 se puede observar la cantidad de valores del conjunto PB que suben, bajan o se mantienen con variaciones mínimas respecto a un dato anterior (según su $PVPR$); y en la tabla 3 se especifican los índices porcentuales definidos para la salida del sistema.

3.1.2. Salida del sistema

Teniendo en cuenta los índices porcentuales y los datos de entrada, se procede a especificar la variable de salida que ha de indicar si el PHBE sube, baja o permanece con variaciones mínimas según el desfase seleccionado.

Dado que existen 3 índices porcentuales para las variaciones del PHBE, se definen 3 clases que identifican la tendencia del precio, como se muestra en la tabla 4.

SALIDA	ÍNDICE PORCENTUAL	DESCRIPCIÓN
$Y = 1$	<-5	El PHBE baja más del 5 %
$Y = 2$	>=-5 $\wedge$ <=5	El PHBE permanece igual o con variaciones mínimas
$Y = 3$	>5	El PHBE sube más del 5 %

Tabla 4. Salidas del sistema.

Por último, se establece el vector de salida del sistema para cada uno de los datos de entrada, como se muestra a continuación:

$$Y_t = \left[Y(t) \right], Y \in \mathcal{M}_{n\times 1}(N) \tag{22}$$

Donde $Y(t)$ es la clase o salida Y para cada t perteneciente a X_t según la ecuación 19.

3.2. METODOLOGÍA PARA PRONÓSTICO CON BASE EN CLASIFICADORES

Para el entrenamiento, prueba y análisis del modelo, se desarrolla una aplicación en Matlab, la cual permite cargar información de las diferentes series de tiempo en los formatos de excel entregados por XM y por el Banco de la República, y a su vez, visualizar las gráficas y tablas correspondientes.

Se dispone de un computador equipado con una tarjeta de video Intel Iris Pro Graphics 6200 de 1536MB, un procesador de 3,3GHz Intel Core i7 de cuatro núcleos y una memoria RAM DDR3 de 16 GB y 1867MHz. Además, se hace uso de la caja de herramientas de computación paralela (PCT, del inglés *Parallel Computing Toolbox*) de Matlab con el fin de incrementar la velocidad de procesamiento de los algoritmos.

Al cargar la información de las diferentes series de tiempo se procede a realizar el ajuste de los datos de entrada como se especifica en la sección 3.1, teniendo en cuenta un desfase de 72 horas para el PHBE, para la Generación Despachada Centralmente y para la Demanda de Energía del SIN; un desfase de 0 horas para el Precio de Escasez y un desfase de 48 horas para el índice COLCAP. Siendo estos valores de desfase, los normalmente conocidos.

Asumiendo que se va a predecir el PHBE del día de mañana y que los últimos datos que se conocen del PHBE, Generación y Demanda, son de dos días atrás, el valor de v_1, v_2 y v_3 conforme a la ecuación 12, para el desfase de estas series respecto a la fecha a predecir, es de 72 horas. Suponiendo también que el último valor del COLCAP es del día de ayer, el valor de v_5, para el desfase del COLCAP, es de menos 48 horas. Y como el Precio de Escasez se conoce con antelación, el valor de v_4 es de 0 horas. De esta forma, y según la ecuación 19, la matriz de datos de entrada está dada por la ecuación 23.

$$X_t = \begin{bmatrix} F_t & PHB_{t-72} & G_{t-72} & DEM_{t-72} & ES_t & COLC_{t-48} \end{bmatrix} \tag{23}$$

Cabe resaltar que también realiza el ajuste del conjunto de datos de salida de acuerdo con los intervalos de fecha y hora de F_t.

Por tanto, la base de datos resultante queda conformada por 113856 registros con información horaria desde el año 2008 hasta el año 2021.

3.2.1. Modelamiento con ANNs

Para el modelamiento con clasificadores univariables se implementa la red ANN, haciendo uso de validación cruzada. De esta forma se divide el conjunto de datos en dos

particiones, una del 70 % para entrenamiento (79698 registros) y la otra del 30 % para prueba (34158 registros), durante 10 iteraciones; creando así 10 subconjuntos diferentes para el entrenamiento de los modelos, y 10 subconjuntos exclusivos para la etapa de prueba o evaluación. Seguidamente se entrenan cinco modelos diferentes para la red ANN como se muestra en la tabla 5. Cabe resaltar que, con el fin de analizar correctamente la información sin ninguna variable sobresaliente a las demás, se normaliza el conjunto de datos.

MODELO	CAPAS OCULTAS	NEURONAS EN CAPAS OCULTAS
1	5	[600, 300, 200, 100, 50]
2	4	[400, 150, 100, 30]
3	4	[200, 50, 10, 5]
4	3	[100, 20, 10]
5	1	50

Tabla 5. Modelos entrenados ANN.

Para el entrenamiento del modelo se configuran los datos de validación con el fin de dar un paro anticipado si tras 10000 épocas el algoritmo tiene un sesgo considerable que le impida mejorar. Por otra parte, la función de activación utilizada para las capas ocultas es la función Sigmoide Tangente Hiperbólica o *tansig* definida en 24, y la función de activación para la capa de salida es la función *Softmax* definida en 25.

$$tansig(i) = \frac{2}{1 + exp(-2i)} - 1 \tag{24}$$

$$Softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{25}$$

La función de rendimiento que permite establecer la regla de detención o paro está definida en 26 y se basa en Entropía Cruzada o *crossentropy*. Esta función permite

comparar el dato predicho por la red con el dato verdadero y calcula la pérdida de entropía cruzada para múltiples clases, mediante la probabilidad de que una observación determinada se asigne a una clase determinada, sumada a todas las clases y observaciones, y normalizada por el número de observaciones. Este proceso se realiza iterando las C posibles clases predichas por la red durante la propagación hacia adelante.

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(T_{i,j}log(X_{i,j}) + (1 - T_{i,j})log(1 - X_{i,j}) \tag{26}$$

Donde $X_{i,j}$ es la respuesta de la red para una clase en particular, $T_{i,j}$ es el valor obtenido en esa clase, y C es el número total de clases.

El proceso de entrenamiento de la red implica ajustar los valores de los pesos y los sesgos de la red para optimizar el rendimiento, según lo define la función 26. Como algoritmo para optimizar esta función se emplea el Gradiente Conjugado Escalado o *trainscg*. Este optimizador es ampliamente utilizado para el reconocimiento de patrones, y aunque sus requisitos de memoria son relativamente pequeños, es mucho más rápido que los algoritmos de descenso de gradiente estándar [51].

Para evaluar el desempeño de los diferentes modelos se utiliza la métrica de Exactitud (*Accuracy*) que indica cuántos valores de tendencia de PHBE son predichos correctamente según el número total de registros. De esta forma se obtienen los resultados por cada modelo, como se muestra en la figura 20.

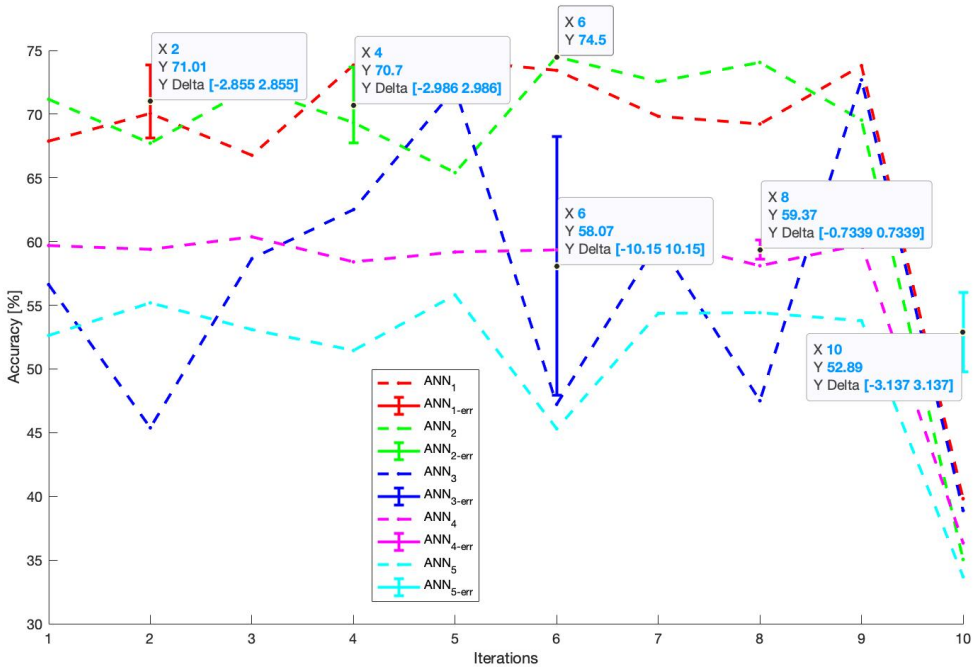


Figura 20. Exactitud modelo ANN. Fuente: autor

Cabe resaltar que, en la última iteración, las etapas de entrenamiento y de prueba se realizan con la serie de datos ordenada por fecha, siendo esta iteración la de menor porcentaje de exactitud. En la figura 20 se evidencia que los modelos con mayor exactitud son el modelo 1 y el 2. Con una media de exactitud del 71.01 % y 70.7 % respectivamente, sin tener en cuenta la última iteración. De la misma forma, el menor nivel de incertidumbre se presenta en el modelo 4 y el mayor nivel en el modelo 3.

Por otro lado se grafican las matrices de confusión de cada modelo, eligiendo la iteración con mayor porcentaje de exactitud, como se muestra en las figuras 21-25.

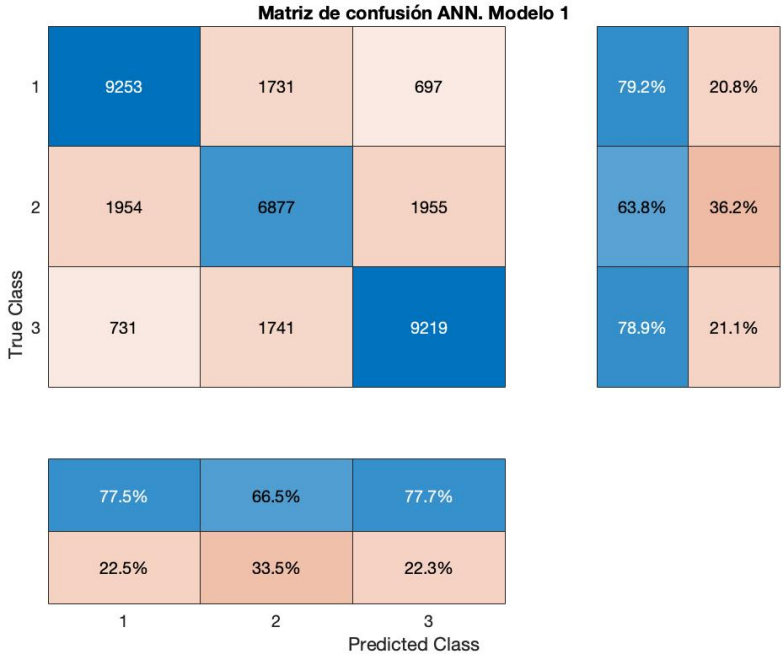


Figura 21. Matriz de confusión Modelo 1. Fuente: autor

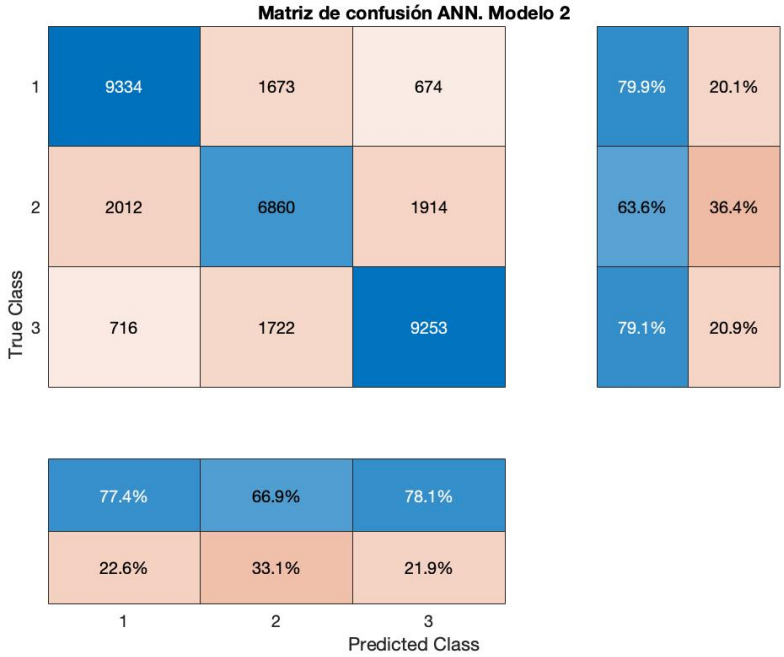


Figura 22. Matriz de confusión Modelo 2. Fuente: autor

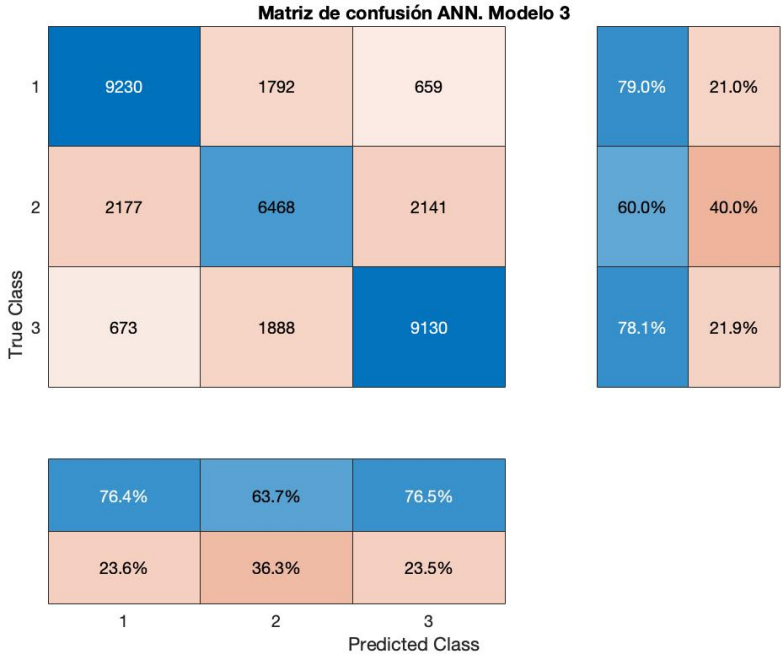


Figura 23. Matriz de confusión Modelo 3. Fuente: autor

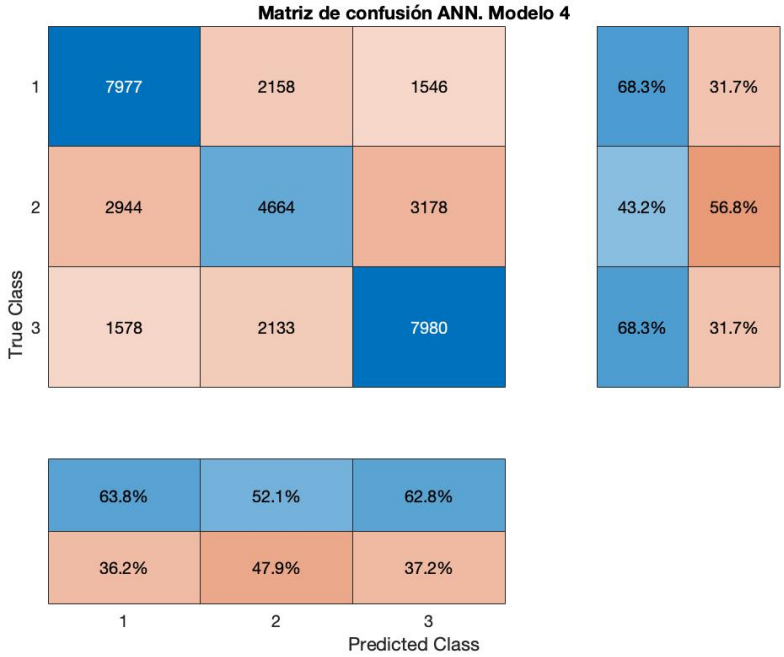


Figura 24. Matriz de confusión Modelo 4. Fuente: autor

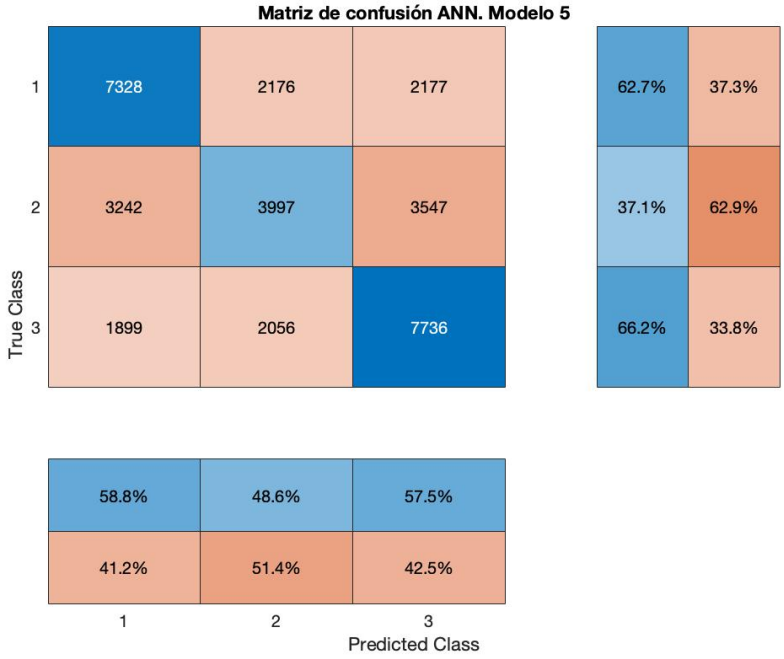


Figura 25. Matriz de confusión Modelo 5. Fuente: autor

La matrices de confusión son otra forma de evaluar el desempeño de los diferentes modelos. En este caso, como se observa en las figuras 21-25, es posible identificar dos métricas; la primera métrica es la Precisión, que indica qué proporción de los datos que son predichos, se clasifican correctamente según el índice porcentual. Y la segunda métrica es la sensibilidad o *Recall*, que indica qué proporción de datos que realmente corresponden a cierto índice porcentual de tendencia son clasificados con dicho índice porcentual.

Finalmente, el promedio por iteración de los tiempos de entrenamiento de los modelos analizados según la tabla 5, se registra en la tabla 6.

	MODELO 1	MODELO 2	MODELO 3	MODELO 4	MODELO 5
TIEMPO (s)	4402.3979	2575.1385	1362.1099	398.2141	12.6088

Tabla 6. Tiempos de entrenamiento ANN.

Al analizar los resultados obtenidos, se observa que el modelo 2 es el que presenta mejor exactitud a menor tiempo de entrenamiento, con un valor de 74.5 %. De acuerdo con su matriz de confusión, la clase 1 tiene una precisión del 79.9 %, la clase 3 una precisión del 79.1 %, y la clase 2 es la que muestra menor precisión con 63.6 %, siendo esta clase la que representa los casos donde el PHBE permanece igual o con variaciones mínimas (con un porcentaje de ocurrencia menor que las otras clases a medida que aumenta el desfase v_1). Por otra parte, la sensibilidad o *Recall* para la clase 1 es de 77.4 %, para la clase 2 de 66.9 %, y para la clase 3 de 78.1 %.

Por lo anterior, se elige el modelo 2 como el modelo más apto para la clasificación.

3.2.2. Selección de características con HMM (HMMFe)

En esta sección se implementan los modelos Markovianos latentes o HMMs con el fin de definir las características relevantes para re-entrenar el modelo elegido en la sección 3.2.1. Se hace uso de validación cruzada dividiendo el conjunto de datos en dos particiones, una del 70 % para entrenamiento (79698 registros) y otra del 30 % para prueba (34158 registros), durante 5 iteraciones. Además se normalizan los datos para evitar variables sobresalientes a las demás.

Para la selección de características relevantes haciendo uso del HMM (HMMFe), se calculan los parámetros del modelo de manera similar a como se plantea en [7] y [44]. Los pasos principales en el cálculo de los parámetros son los siguientes:

1. Se genera la matriz de transición para los estados ocultos, de acuerdo con la ecuación 27, donde cada elemento representa el comportamiento de las observaciones en el tiempo según los estados del sistema.

2. Se calcula la distribución inicial de estados que representa la probabilidad para

cada uno de los estados iniciales o para el conjunto predefinido de clases, como se muestra en la ecuación 28.

3. Se aplica el método *K-Means* con el fin de encontrar los *clusters* del conjunto de entrenamiento y así poder discretizar o cuantizar los datos de entrada, por cada característica o columna de la matriz X_t definida en 19. Para este trabajo se define una cantidad de *clusters* igual al valor redondeado de la desviación estándar de la columna de datos correspondiente, y se hace uso de la ecuación 30 con el fin de discretizar los datos de acuerdo con los *clusters* encontrados.

4. Se calcula la matriz de probabilidades de emisión, como se define en 29, la cual representa las probabilidades de que a una observación le corresponda un estado específico. Esta matriz se calcula haciendo uso del conjunto de datos de entrenamiento cuantizados en el paso 3, por cada característica.

$$A = \{a_{ij}\},\ a_{ij} = P(s_j|s_i),\ i \in N, j \in N$$
$$A = \frac{\sum S_i \rightarrow S_j}{\sum_{t=1}^{T} s^t = S_i},\ s^t = S_i, s^{t+1} = S_j \tag{27}$$

Donde $S \in \{S_n\}$, y $\{S_n\}$ es el conjunto de la lista única de clases para cada salida. Por otro lado, $s^t \in S\ \forall\ t = 1, 2.., T$, siendo T el tamaño del conjunto de datos de entrenamiento. Y, $\{S_i \rightarrow S_j\}$ representa los estados futuros que le corresponden a los estados previos.

$$\pi = \frac{1}{\sum S_i},\ i \in N \tag{28}$$

$$B = \{b_i(k)\},\ b_i(k) = P(o_k|s_i),\ i \in N$$

$$B = \frac{\sum S_i \rightarrow O_i}{\sum_{t=1}^{T} s^t = S_i},\ s^t = S_i \tag{29}$$

Donde $\{S_i \rightarrow O_i\}$ representa las muestras u observaciones que le corresponden al estado S_i.

$$Z(k) = min(\sum_{j=1}^{m} e_{kj}),\ E = \{e_{kj}\} \in \mathcal{M}_{n\times m},\ k = 1, ..., n \tag{30}$$

Donde E es la distancia euclídea entre los *clusters* de los datos de entrenamiento (C_{train}) y el conjunto de observaciones (O), como se muestra en la ecuación 31. Y, en consecuencia, $Z(k)$ es el vector de la mínima distancia euclídea entre estos datos.

$$E = [C_{train} - O(i)]^2,\ C_{train} \in \mathcal{M}_{k\times m},\ O \in \mathcal{M}_{n\times m},\ i = 1, ..., n \tag{31}$$

Luego de definir los parámetros del modelo $\mu = (A, B, \pi)$, para cada característica del conjunto de datos, se calcula la longitud mínima del LogLik de cada uno de los modelos resultantes haciendo uso de la secuencia de datos correspondiente, misma secuencia que fue cuantizada para la definición de la matriz de probabilidades de emisión. Seguidamente, se normalizan dichos valores de LogLik, entre 0 y 1, con el fin de utilizarlos para definir el peso y el índice de las características en el conjunto de datos. El vector resultante es reflejado y desplazado en 1, para luego ser ordenado según el peso definido por el LogLik con el fin de establecer el orden de relevancia de las características.

Para evaluar el método de selección de características HMMFe, se realiza una com-

paración con los métodos de ReliefF y PCA, en cada iteración. De esta manera se ordenan las 22 características de X_t de acuerdo con el peso de relevancia (normalizado) correspondiente, como se observa en la figura 26.

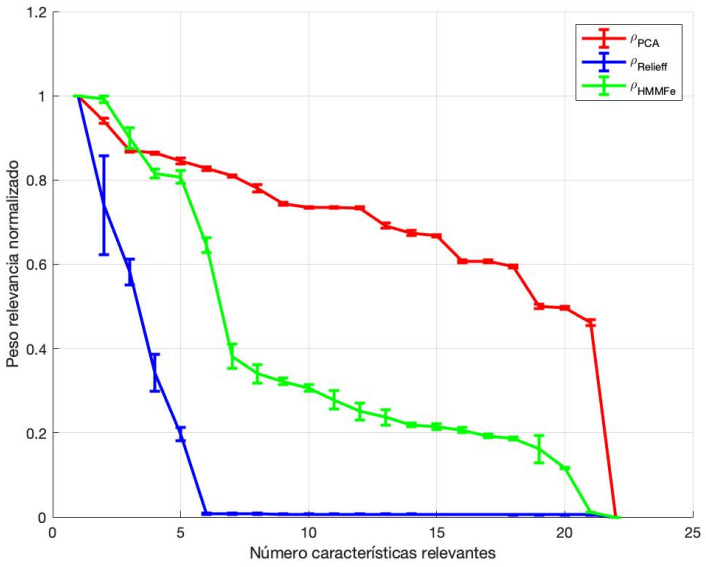


Figura 26. Relevancia de características PHBE. Fuente: autor

Para las 5 iteraciones, el peso de relevancia por número de características no varía significativamente, sin embargo, se presenta mayor grado de variación en el algoritmo de ReliefF dada su desviación estándar, seguido del método HMMFe. Adicionalmente, se observa que para el caso de ReliefF existen 5 características con peso relevante, para el caso de PCA exiten 21 características, y para HMMFe existen 19 o 20. Las demás características pueden ser contraproducentes en la predicción.

El promedio de los tiempos, para el cálculo del peso por relevancia de las características, se registra en la tabla 7.

	PCA	HMMFe	ReliefF
TIEMPO (s)	0.019649	136.5025	518.5538

Tabla 7. Tiempos de selección de características.

Posteriormente se entrena el modelo elegido en la sección 3.2.1 para cada número de características relevantes según los métodos de selección. Se utiliza la métrica *Accuracy*, con el fin de evaluar el desempeño de la clasificación, como se muestra en la figura 27.

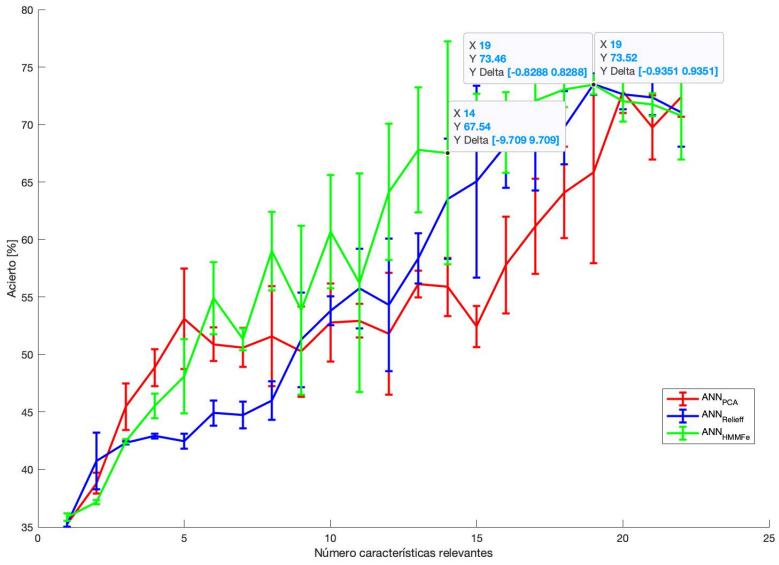


Figura 27. Exactitud del modelo ANN según características relevantes. Fuente: autor

Se evidencia que el método de selección de características HMMFe presenta mayor exactitud que los demás métodos, para los modelos entre 6 y 18 características relevantes. Asimismo, el método de ReliefF presenta una media de exactitud del 73.52 % para 19 características, mientras que HMMFe presenta una media del 73.46 % y un nivel de incertidumbre ligeramente menor al de ReliefF en este punto. En cuanto a PCA se observa una media de exactitud del 72.88 % con 20 características. Siendo estos puntos los de mayor acierto. Por otra parte, el punto que presenta mayor grado de incertidumbre con mayor media de exactitud es HMMFe, a 14 características, siendo su media de exactitud del 67.54 %. Cabe resaltar que ReliefF requiere mayor tiempo de ejecución que los demás métodos, y PCA es el método que presenta menor exigencia computacional.

El orden por relevancia de características, que es presentado de forma más común con

el método de selección HMMFe, de mayor a menor relevancia, es el siguiente: TG_t, HG_t, PB_t, ED_t, C_t, E_t, horas, días, meses, años, y números del día de la semana. Las características de SG_t, EG_t y OG_t manifiestan poca relevancia debido a que la mayoría de datos se encuentran en 0, y por tanto, no son tenidas en cuenta para el entrenamiento de los modelos analizados, sin embargo, influyen en las demás características al momento de la normalización de los datos.

3.3. VALIDACIÓN DE LA METODOLOGÍA

En esta investigación se propone implementar la selección de características HMMFe, descrita en la sección 3.2.2, en el modelo ANN descrito en la sección 3.2.1. Consiguientemente, el diagrama que identifica el modelo propuesto se puede apreciar en la figura 28.

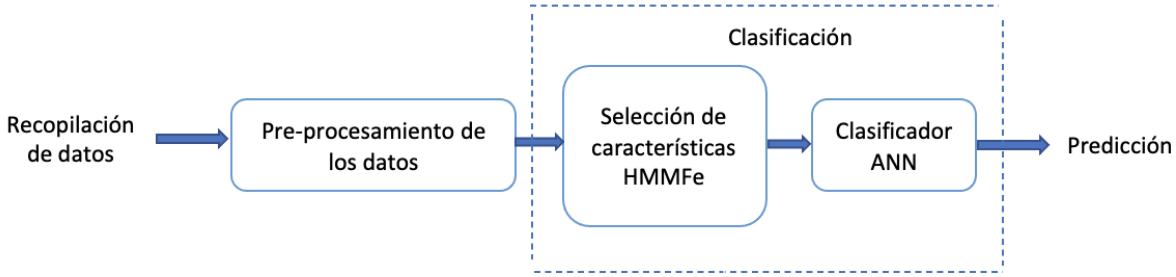


Figura 28. Diagrama del modelo propuesto. Fuente: autor

Para el entrenamiento del modelo, inicialmente se hace uso de validación cruzada y se normalizan los datos de acuerdo con la partición, como se especifica en la sección 3.2.1. Seguidamente se realiza el cálculo de características relevantes HMMFe utilizando el conjunto de datos normalizado para evitar variables sobresalientes a las demás, cabe resaltar que el orden de las características relevantes no varía de forma significativa sin importar la partición del conjunto de datos tomada.

Se realiza el entrenamiento para 19 características y para 14 características relevantes, dando como resultado lo que se evidencia en las figuras 29 y 30.

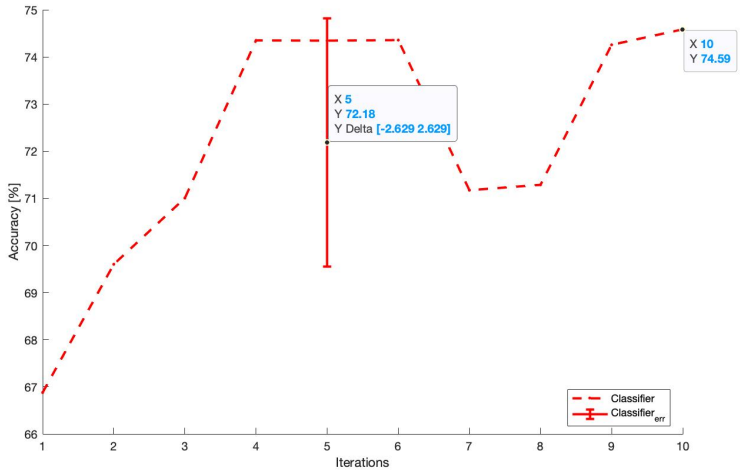


Figura 29. Exactitud modelo 19 características. Fuente: autor

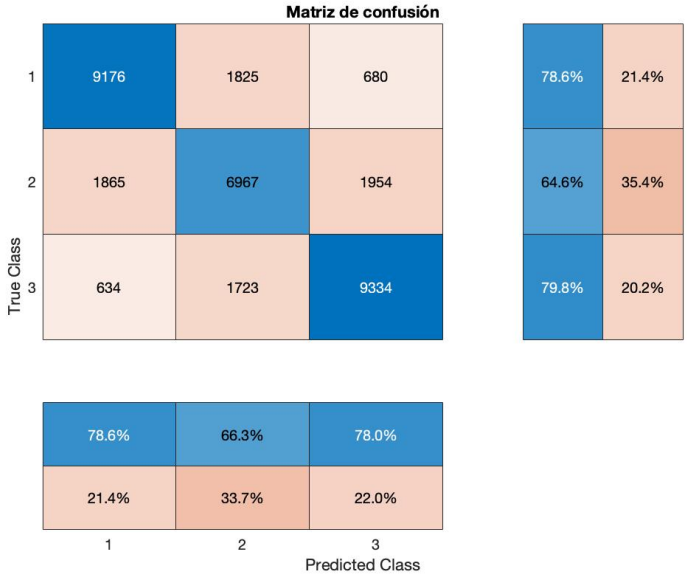


Figura 30. Matriz de confusión modelo 19 características. Fuente: autor

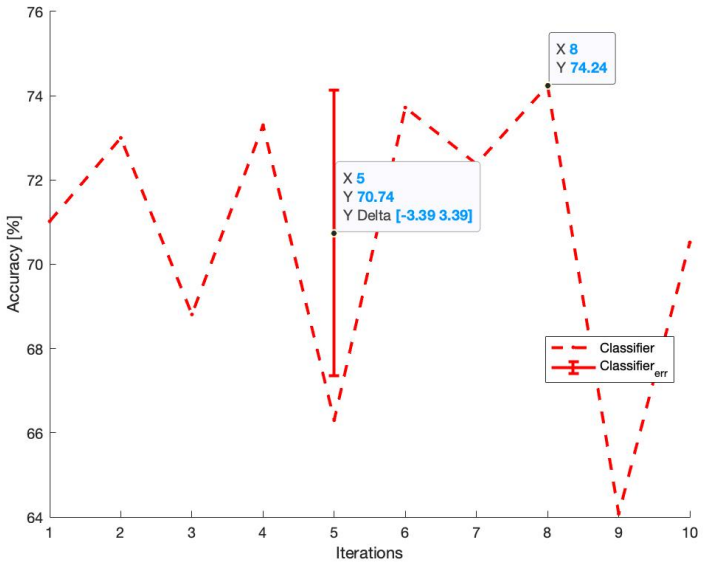


Figura 31. Exactitud modelo 14 características. Fuente: autor

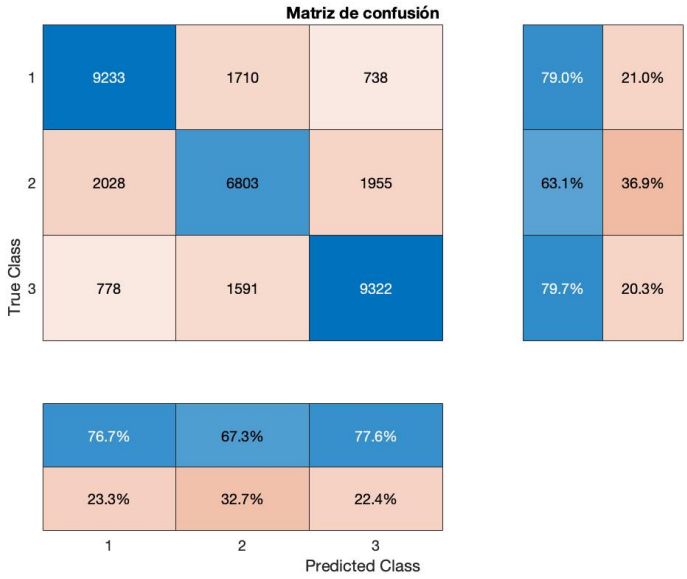


Figura 32. Matriz de confusión modelo 14 características. Fuente: autor

Según la figura 29, el modelo con 19 características relevantes presenta mayor exactitud y menor desviación estándar que el modelo seleccionado en la sección 3.2.1, con una

media de exactitud del 72.18 %. El porcentaje de mayor exactitud es observado en la última iteración con 74.6 %.

La matriz de confusión presentada en la figura 30, para la última iteración, evidencia que la clase 1 tiene una precisión del 78.6 %, la clase 3 una precisión del 79.8 %, y la clase 2 es la que muestra menor precisión con 64.6 %. Por otra parte, la sensibilidad o *Recall* para la clase 1 es de 78.6 %, para la clase 2 de 66.3 %, y para la clase 3 de 78 %.

El tiempo promedio de ejecución del algoritmo, por iteración, es de 2026.632 segundos.

Por otra parte, según la figura 31, el modelo con 14 características relevantes presenta una media de exactitud del 70.74 %. El porcentaje de mayor exactitud es observado en la octava iteración con 74.24 %. Sin embargo, la desviación estándar es considerable.

La matriz de confusión presentada en la figura 32, para la última iteración, evidencia que la clase 1 tiene una precisión del 79.0 %, la clase 3 una precisión del 79.7 %, y la clase 2 es la que muestra menor precisión con 63.1 %. Por otra parte, la sensibilidad o *Recall* para la clase 1 es de 76.7 %, para la clase 2 de 67.3 %, y para la clase 3 de 77.6 %.

El tiempo promedio de ejecución del algoritmo, por iteración, es de 1712.07 segundos.

En la tabla 8 se realiza una comparación de los 3 modelos seleccionados, con mejor desempeño. El modelo elegido en la sección 3.2.1 y los modelos analizados en la presente sección.

Modelo inicial			
Tiempo promedio [s]	Accuracy [%]	Precisión [%]	Recall [%]
2575.1385	74.5	74.2	74.14
Modelo HMMFe 19 características			
2026.632	74.6	74.34	74.3
Modelo HMMFe 14 características			
1712.07	74.24	73.94	73.87

Tabla 8. Tabla de comparación de los modelos

Finalmente se selecciona el modelo HMMFe de 19 características para la predicción del PHBE con el fin de ser utilizado en la aplicación desarrollada.

3.4. CONCLUSIONES Y RECOMENDACIONES

3.4.1. Conclusiones

Al realizar el análisis de la información documentada en el estado del arte se confirma que los clasificadores ANNs son una herramienta potente y flexible para pronosticar la tendencia del PHBE, además, se pueden combinar de muchas formas con los HMMs para mejorar su desempeño.

Entre las características y parámetros adaptativos al mercado eléctrico, que se definen en este trabajo, se evidencia que las variables que permiten identificar la matriz de generación, el PHBE y la variable de demanda energética a nivel nacional, son las variables con mayor relevancia en las características. Por otra parte, el uso de índices porcentuales en la predicción de la tendencia del PHBE es posible y es un indicador que aporta mayor información en la predicción, en contraste con la sola tendencia de subida o de bajada del precio, debido a que permite identificar la diferencia porcentual respecto al precio anterior con la que tiende a cambiar el precio siguiente. De esta forma, los analistas y otros agentes del mercado pueden tener una idea más acertada para la toma de decisiones.

Además, se hace uso de dos fuentes de información, una de ellas es la información de energía suministrada por la empresa XM, y la otra es la información correspondiente al índice COLCAP proporcionada por el Banco de la República. Evidenciando que el uso de dos o más fuentes de información, como la correspondiente al índice COLCAP, es favorable para la predicción del PHBE dado el orden de relevancia que presentan

las características. Siendo el índice COLCAP una variable con mayor relevancia que el Precio de Escasez.

Por consiguiente, se desarrolla una metodología que permite predecir, a corto plazo, la tendencia con índices porcentuales del PHBE en el mercado eléctrico colombiano, haciendo uso de clasificadores con parámetros adaptativos. Esta metodología presenta un aporte principal para el uso de clasificadores ANNs con modelos Markovianos latentes, dicho aporte se basa en la implementación del HMM como método para selección de características relevantes.

En la implementación de esta metodología se comprueba que es posible reducir la complejidad del algoritmo de clasificación haciendo uso del método de selección de características, denominado HMMFe, con mejor exactitud que otros procedimientos como PCA y ReliefF, y con menor tiempo de ejecución que el algoritmo de ReliefF. Además, se verifica que las características y parámetros definidos para la predicción de la tendencia del PHBE, como la demanda y la generación por diferentes fuentes de energía, son variables importantes tanto para la clasificación como para la sostenibilidad del modelo, ya que indican el comportamiento del mercado eléctrico y sus posibles modificaciones en el futuro.

3.4.2. Recomendaciones

Cuando se realiza el entrenamiento del clasificador haciendo uso de algún método de selección de características relevantes, las características de generación de energía eólica, solar y de otras fuentes de generación poco prominentes, obtienen bajo peso de relevancia. Por tanto, no son consideradas en el modelo final. Como recomendación, es necesario re-entrenar el modelo a medida que estas variables vayan cambiando en el mercado, dado que pueden llegar a tener un alto peso de relevancia, por el momento

no proporcionan ninguna información en la predicción del PHBE ya que la mayoría de estos datos son nulos y pueden ser poco favorables para la clasificación.

3.4.3. Trabajos derivados

Como parte del desarrollo del presente proyecto se presentan a consideración los siguientes productos que pueden servir como referencia para diversos trabajos investigativos permitiendo la reproducibilidad de la metodología presentada:

1. **Nombre del producto**: Ponencia.

   Descripción: Invitación en calidad de conferencista al I Congreso Internacional de Electromecánica, Mecánica y Automatización.

   País: Ecuador.

   Estado del producto: Documento aceptado y aprobado. Congreso programado para los días 26, 27 y 28 de abril del año en curso.

2. **Nombre del producto**: Ponencia.

   Descripción: XXIII Congreso Internacional de Mantenimiento y Gestión de Activos.

   País: Colombia.

   Estado del producto: Resumen aceptado. Conferencia programada para el mes de mayo del año en curso.

3. **Nombre del producto**: Aplicación de software.

   Descripción: Aplicación desarrollada en Matlab, con la cual se puede cargar información nueva en las bases de datos (mediante los formatos de excel correspondientes), realizar el ajuste y el estudio de la información, el análisis de los clasificadores así como el análisis de los métodos de selección de características,

el entrenamiento del modelo y la predicción final.

Estado del producto: Pendiente de registro en la Dirección Nacional de Derechos de Autor y de la cesión de derechos patrimoniales a la universidad Tecnológica de Pereira.

3.4.4. Trabajos futuros

1. Incluir en el modelo propuesto otras variables económicas y otras fuentes de información con el fin de mejorar la predicción del PHBE.

2. Realizar un análisis del comportamiento de la metodología con otros tipos de clasificadores como los SVMs, utilizando el método de selección de características desarrollado.

BIBLIOGRAFÍA

[1] LIU, Hui y LONG, Zhihao. An improved deep learning model for predicting stock market price time series. En: Digital Signal Processing: A Review Journal, tomo 102, 2020. ISSN 10512004. 1.1

[2] LONG, Wen; LU, Zhichen y CUI, Lingxiao. Deep learning-based feature engineering for stock price movement prediction. En: Knowledge-Based Systems, tomo 164, 2019, págs. 163–173. ISSN 09507051. 1.1, 2.1

[3] ZHONG, Xiao y ENKE, David. Forecasting daily stock market return using dimensionality reduction. En: Expert Systems with Applications, tomo 67, 2017, págs. 126–139. ISSN 09574174. 1.1, 2.1, 2.2.1, 2.3

[4] HOSEINZADE, Ehsan y HARATIZADEH, Saman. CNNpred: CNN-based stock market prediction using a diverse set of variables. En: Expert Systems with Applications, tomo 129, 2019, págs. 273–285. ISSN 09574174. 1.1, 1.2, 2.1

[5] MAQSOOD, Haider, *et al.* A local and global event sentiment based efficient stock exchange forecasting using deep learning. En: International Journal of Information Management, tomo 50, Nº December 2018, 2020, págs. 432–451. ISSN 02684012. 1.1, 2.1, 2.2.1

[6] CHOURMOUZIADIS, Konstandinos y CHATZOGLOU, Prodromos D. An intelligent short term stock trading fuzzy system for assisting investors in portfolio management. En: Expert Systems with Applications, tomo 43, 2016, págs. 298–311. ISSN 09574174. 1.1

[7] GE, Yidi, *et al.* Beyond negative and positive: Exploring the effects of emotions in social media during the stock market crash. En: Information Processing and

Management, tomo 57, Nº 4, 2020, pág. 102218. ISSN 03064573. 1.1, 1.2, 2.1, 2.2.1, 2.4, 3.2.2

[8] CHEN, Yingjun y HAO, Yongtao. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. En: Expert Systems with Applications, tomo 80, 2017, págs. 340–355. ISSN 09574174. 1.1, 1.2

[9] HENRIQUE, Bruno Miranda; SOBREIRO, Vinicius Amorim y KIMURA, Herbert. Literature review: Machine learning techniques applied to financial market prediction. En: Expert Systems with Applications, tomo 124, 2019, págs. 226–251. ISSN 09574174. 1.1, 1.2, 2.1, 2.6.1

[10] CHONG, Eunsuk; HAN, Chulwoo y PARK, Frank C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. En: Expert Systems with Applications, tomo 83, 2017, págs. 187–205. ISSN 09574174. 1.1, 1.2, 2.1

[11] WENG, Bin; AHMED, Mohamed A. y MEGAHED, Fadel M. Stock market one-day ahead movement prediction using disparate data sources. En: Expert Systems with Applications, tomo 79, 2017, págs. 153–163. ISSN 09574174. 1.1, 1.2

[12] PADHIARY, Pratap Kishore y MISHRA, Ambika Prasad. Development of Improved Artificial Neural Network Model for Stock Market Prediction. En: International Journal of Engineering Science and Technology, tomo 3, Nº 2, 2011, págs. 1576–1581. ISSN 0975-5462. 1.1, 1.2, 2.1

[13] LAHMIRI, Salim y BEKIROS, Stelios. Cryptocurrency forecasting with deep learning chaotic neural networks. En: Chaos, Solitons and Fractals, tomo 118, 2019, págs. 35–40. ISSN 09600779. 1.1

[14] SINGH, Ritika y SRIVASTAVA, Shashi. Stock prediction using deep learning. En: Multimedia Tools and Applications, tomo 76, Nº 18, 2017, págs. 18569–18584. ISSN 15737721. 1.1, 1.2, 2.1

[15] HASSAN, Md Rafiul. A combination of hidden Markov model and fuzzy model for stock market forecasting. En: Neurocomputing, tomo 72, Nº 16-18, 2009, págs. 3439–3446. ISSN 09252312. 1.1, 1.2

[16] CHENG, Ching Hsue y YANG, Jun He. Fuzzy time-series model based on rough set rule induction for forecasting stock price. En: Neurocomputing, tomo 302, 2018, págs. 33–45. ISSN 18728286. 1.1, 2.1

[17] BASAK, Suryoday, *et al.* Predicting the direction of stock market prices using tree-based classifiers. En: North American Journal of Economics and Finance, tomo 47, Nº December 2017, 2019, págs. 552–567. ISSN 10629408. 1.1, 1.2

[18] CHANDRIKA, P. V.; VISALAKSHMI, K. y SAKTHI SRINIVASAN, K. Application of Hidden Markov Models in Stock Trading. En: 2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020, 2020, págs. 1144–1147. 1.1, 1.2, 2.1

[19] GUPTA, Aditya y DHINGRA, Bhuwan. Stock market prediction using Hidden Markov Models. En: 2012 Students Conference on Engineering and Systems, SCES 2012, 2012, págs. 1–4. 1.1, 1.2, 2.1, 2.4

[20] COMISIÓN DE REGULACIÓN DE ENERGÍA Y GAS. D-066- AUTOGENERACIÓN A PEQUEÑA ESCALA Y GENERACIÓN DISTRIBUIDA, 2017. 1.2

[21] MOHSENIAN-RAD, Amir Hamed y LEON-GARCIA, Alberto. Optimal residential load control with price prediction in real-time electricity pricing environments.

En: IEEE Transactions on Smart Grid, tomo 1, Nº 2, 2010, págs. 120–133. ISSN 19493053. 1.2

[22] ANBAZHAGAN, S. y KUMARAPPAN, N. Day-ahead deregulated electricity market price forecasting using neural network input featured by DCT. En: Energy Conversion and Management, tomo 78, Nº 4, 2014, págs. 711–719. ISSN 01968904. 1.2, 2.1

[23] WANG, Feng; LIU, Ling y DOU, Chenxiao. Stock market volatility prediction: A service-oriented multi-kernel learning approach. En: Proceedings - 2012 IEEE 9th International Conference on Services Computing, SCC 2012, tomo d, 2012, págs. 49–56. 1.2, 2.1

[24] YU, Pengfei y YAN, Xuesong. Stock price prediction based on deep neural networks. En: Neural Computing and Applications, tomo 32, Nº 6, 2020, págs. 1609–1628. ISSN 14333058. 2.1

[25] NABIPOUR, Mojtaba, *et al.* Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; A Comparative Analysis. En: IEEE Access, tomo 8, 2020, págs. 150199–150212. ISSN 21693536. 2.1, 2.3

[26] IDREES, Sheikh Mohammad; ALAM, M. Afshar y AGARWAL, Parul. A Prediction Approach for Stock Market Volatility Based on Time Series Data. En: IEEE Access, tomo 7, 2019, págs. 17287–17298. ISSN 21693536. 2.2

[27] RUIZ, L. G.B., *et al.* Energy consumption forecasting based on Elman neural networks with evolutive optimization. En: Expert Systems with Applications, tomo 92, 2018, págs. 380–389. ISSN 09574174. 2.2

[28] XM. Descripción del sistema eléctrico Colombiano ¿A qué se dedica XM? ¿Qué tipo de operaciones se realizan en la Bolsa de Energía? ¿Cómo funciona el mercado mayorista eléctrico en Colombia? URL http://www.xm.com.co/Paginas/Mercado-de-energia/descripcion-del-sistema-electrico-colombiano.aspx. [Web; Consultado el 13/01/2021]. 2.2.1

[29] XM. Precio de Bolsa y Escasez. URL https://www.xm.com.co/Paginas/Mercado-de-energia/precio-de-bolsa-y-escasez.aspx. [Web; Consultado el 13/01/2021]. 2.2.1

[30] BHATTACHARYYA, Subhes. Energy Economics (Manual). Springer, 2011. ISBN 9780857292674, 723 págs. 2.2.1

[31] LI, Xiaodong; WU, Pangjing y WANG, Wenpeng. Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. En: Information Processing and Management, tomo 57, Nº 5, 2020, pág. 102212. ISSN 03064573. 2.2.1

[32] XM. Histórico de Precios. URL http://portalbissrs.xm.com.co/trpr/Paginas/Historicos/Historicos.aspx. [Web; Consultado el 14/01/2021]. 2.2.1

[33] XM. Precio Promedio de Bolsa. URL http://portalbissrs.xm.com.co/trpr/Paginas/Precios/preprobol.aspx. [Web; Consultado el 14/01/2021]. 1

[34] XM. Histórico de Demanda Energética. URL http://portalbissrs.xm.com.co/dmnd/Paginas/Historicos/Historicos.aspx. [Web; Consultado el 14/01/2021]. 2.2.2

[35] XM. Históricos de demanda de energía en Colombia. URL https://www.xm.com.co/Paginas/Consumo/historico-de-demanda.aspx. [Web; Consultado el 25/01/2021]. 2.2.2, 5

[36] BANCO DE LA REPÚBLICA. COLCAP. URL https://www.banrep.gov.co/es/estadisticas/mercado-accionario. [Web; Consultado el 18/01/2021]. 2.2.3, 7

[37] ANDRADE, FABRICIO CORONEL. Economista, especialista en Finanzas, aspirante a la MAF. Profesional en Comercio Internacional, especialista en Finanzas, aspirante a la MAF. En: , 2019. 2.2.3

[38] XM. Tipos de generación de energía en Colombia. URL https://www.xm.com.co/Paginas/Generacion/tipos.aspx. [Web; Consultado el 22/01/2021]. 2.2.4

[39] XM. Históricos de generación de energía en Colombia. URL http://portalbissrs.xm.com.co/oferta/Paginas/Historicos/Historicos.aspx. [Web; Consultado el 21/01/2021]. 2.2.4

[40] IVAN NUNES DA SILVA, DANILO HERNANE SPATTI, ROGERIO ANDRADE FLAUZINO , LUISA HELENA BARTOCCI LIBONI, Silas Franco dos Reis Alves. Artificial Neural Networks A Practical Course. Springer, 2017. ISBN 9783319431611, 307 págs. 2.3, 2.3

[41] NABIPOUR, M., *et al.* Deep learning for stock market prediction. En: Entropy, tomo 22, Nº 8, 2020. ISSN 10994300. 10, 11

[42] ZUCCHINI, Walter; MACDONALD, Iain L. y BIELEFELD, Roland Langrock. Hidden Markov Models for Time Series: An Introduction using R. General Editors, 2017, 278 págs. 2.4, 2.4, 12

[43] VITERBI, Andrew. Viterbi algorithm. En: Scholarpedia, tomo 4, Nº 1, 2009, pág. 6246. ISSN 1941-6016. 2.4, 2.4

[44] ESMAEL, Bilal, *et al.* Improving time series classification using Hidden Markov Models. En: Proceedings of the 2012 12th International Conference on Hybrid Intelligent Systems, HIS 2012, 2012, págs. 502–507. 2.4, 3.2.2

[45] JAIN, Anil K. Data clustering: 50 years beyond K-means. En: Pattern Recognition Letters, tomo 31, Nº 8, 2010, págs. 651–666. ISSN 01678655. 2.5, 13

[46] YU, Huanhuan; CHEN, Rongda y ZHANG, Guoping. A SVM stock selection model within PCA. En: Procedia Computer Science, tomo 31, 2014, págs. 406–412. ISSN 18770509. 2.6.1

[47] GOLLOU, Abbas Rahimi y GHADIMI, Noradin. A new feature selection and hybrid forecast engine for day-ahead price forecasting of electricity markets. En: Journal of Intelligent and Fuzzy Systems, tomo 32, Nº 6, 2017, págs. 4031–4045. ISSN 18758967. 2.6.2

[48] ROBNIK SIKONJA MARKOROBNIK, Marko y FRIUNI-LJSI IGOR KONONENKO IGORKONONENKO. Theoretical and Empirical Analysis of ReliefF and RReliefF. En: Machine Learning, tomo 53, 2003, págs. 23–69. 2.6.2

[49] GONZALEZ, Ligdi. Amazon Machine Learning Guía para desarrolladores. En: , 2018. 2.7

[50] BVC. Horarios de negociación de la Bolsa de Valores Colombiana. URL https://www.bvc.com.co/pps/tibco/portalbvc/Home/Mercados/horariosDeNegociacion?action=dummy. [Web; Consultado el 21/01/2021]. 3.1

[51] XM. Multilayer NN. URL https://la.mathworks.com/help/deeplearning/ug/train-and-apply-multilayer-neural-networks.html. [Web; Consultado el 08/02/2021]. 3.2.1